

WHEN THE BRAIN TIMES ITSELF:

Neural and Behavioral Mechanisms Underlying

Temporal Error Monitoring

"Gdy Mózg Ocenia Swój Subiektywny Czas: Neuronalne i Behawioralne Mechanizmy Leżące u Podstaw Monitorowania Błędów Czasowych"

"Beyin Kendi İçsel Zamanını Ölçerken: Zamansal Hata İzlemenin Temelindeki Nöral ve Davranışsal Mekanizmalar"

Sena Nur BILGIN

(M.A., mgr)

Doctoral thesis under the supervision of:

Dr. hab. Tadeusz Władysław Kononowicz

Institute of Psychology, Polish Academy of Sciences

&

Graduate School for Social Research (GSSR)

Warsaw, 2026

This thesis is submitted to an institution in Poland, a country whose commitment to women in science is perhaps best embodied by Maria Skłodowska-Curie, the first woman to win a Nobel Prize, and is dedicated to Mustafa Kemal Atatürk, the foundational leader of the Turkish Republic, who opened the doors of science and university education to women. Two nations, two paths, one conviction: that women belong in science.

"If one day, my words are against science, choose science."

"Eğer bir gün benim sözlerim bilimle ters düşerse, bilimi seçin."

"Jeśli pewnego dnia moje słowa będą sprzeczne z nauką, wybierzcie naukę."

— Mustafa Kemal Atatürk

Abstract

Metacognition, the capacity to monitor and evaluate one's own cognitive states, encompasses a broad class of self-monitoring abilities. Temporal error monitoring (TEM) is one such ability, defined as the metacognitive process by which individuals assess the accuracy of self-generated temporal intervals without external feedback. As a tractable instance of metacognition, TEM offers a window into the computational and neural substrates of self-monitoring. Motor action timing serves as the experimental framework for investigating TEM: it generates a precise, internally-accessible error signal, the deviation between a produced interval and a target, and withholds external feedback, isolating metacognitive evaluation from externally-driven correction. This thesis presents a systematic investigation of TEM across complementary levels of analysis, integrating cross-species behavioral paradigms, motor-demand manipulations, and single-trial neural decoding via a transformer-based EEG architecture.

A first question concerned whether TEM is a conserved capacity across species and whether it is supported by shared or divergent strategies. Rats and humans were tested under matched conditions, with aligned trial structures, reward contingencies, and task rules, and were required to select the correct reward option based on the magnitude of their timing error relative to a target interval. Both species demonstrated reliable trial-by-trial error monitoring, establishing TEM as a shared capacity. The information supporting this monitoring, however, differed: rats relied more heavily on current-trial evidence, whereas humans drew more on accumulated trial history and learned reward rules. Critically, when task rule comprehension was equated across species, differences in TEM performance diminished, indicating that the shared monitoring capacity is shaped by strategy selection and rule use rather than by fundamental differences in timing ability.

TEM's conservation across species raised a further question: what information grounds metacognitive judgments about motor action timing in humans? Two competing hypotheses were formulated. Under the Clock Hypothesis, TEM precision decreases as clock signal noise increases, because additional motor variance disrupts the readout of the internal temporal representation. Under the Motor Hypothesis, motor variability improves TEM precision by providing a reference point between intended and actual

motor actions. To dissociate these accounts, motor signal variance was experimentally manipulated by contrasting a button-press condition with a joystick-movement condition. Increased motor demand selectively impaired metacognitive precision without affecting motor action timing accuracy, supporting the Clock Hypothesis: individuals evaluate internal temporal representations rather than motor output signals. This aligns with a temporal metacognition account in which higher-order confidence judgments reflect a metacognitive readout of timing variability rather than motor execution.

Yet knowing that TEM draws on internal timing representations leaves unresolved whether metacognitive evaluation merely reads out the signals governing interval production or constitutes a neurally distinct process. The third study addressed this directly by decoding PCA-optimized single-trial EEG data using a Vision Transformer capturing theta (θ; 3–7 Hz), alpha (α; 8–12 Hz), and beta (β; 13–30 Hz) dynamics across both the motor action timing and TEM decision phases. Motor action timing accuracy was decodable from any single frequency band, consistent with independent frequency-specific oscillatory codes. Metacognitive inference, the alignment between objective timing performance and subjective error evaluation, by contrast, required the joint contribution of all three bands during ongoing motor action timing, with no single band sufficient alone. Furthermore, individuals whose pre-termination neural activity more reliably encoded metacognitive states exhibited greater TEM precision. Together, these findings establish metacognitive evaluation as a neurally distinct process, supported by higher-order multi-band integration.

Across the three studies, internal timing representations formed the core substrate of TEM, shaped by task rules, motor demands, and reward expectations. Impaired TEM precision under higher motor demands supports a clock-based account, while the dissociation between motor timing codes and multi-band metacognitive inference indicates that TEM relies on higher-order neural mechanisms distinct from performance itself.

Keywords: metacognition, temporal error monitoring (TEM), motor action timing, interval production, EEG decoding, transformer architecture.

Streszczenie

Metapoznanie, rozumiane jako zdolność monitorowania i oceny własnych procesów poznawczych, obejmuje szeroki zakres zdolności samomonitorowania. Monitorowanie błędów temporalnych (TEM) jest jedną z takich zdolności, definiowaną jako proces metapoznawczy, w którym jednostki oceniają dokładność samodzielnie generowanych interwałów czasowych bez zewnętrznej informacji zwrotnej. Jako podatna na badania instancja metapoznania, TEM oferuje wgląd w obliczeniowe i neuronalne podłoże samomonitorowania. Czasowanie działań motorycznych służy jako ramy eksperymentalne do badania TEM: generuje precyzyjny, wewnętrznie dostępny sygnał błędu, odchylenie między wyprodukowanym interwałem a docelowym i wstrzymuje zewnętrzną informację zwrotną, izolując ocenę metapoznawczą od zewnętrznie napędzanych korekt. Niniejsza praca przedstawia wielopoziomowe badanie TEM, łącząc paradygmaty behawioralne obejmujące wiele gatunków, manipulacje wymaganiami motorycznymi oraz dekodowanie danych EEG na poziomie pojedynczych prób przy użyciu architektury opartej na transformerze.

Pierwsze pytanie dotyczyło tego, czy TEM jest zachowaną zdolnością u różnych gatunków i czy jest wspierana przez wspólne czy rozbieżne strategie. Szczury i ludzie byli testowani w dopasowanych warunkach, z wyrównanymi strukturami prób, warunkami nagradzania i regułami zadania i musieli wybrać właściwą opcję nagrody na podstawie wielkości błędu czasowania względem docelowego interwału. Oba gatunki wykazały wiarygodne monitorowanie błędów próba po próbie, ustanawiając TEM jako wspólną zdolność. Informacje wspierające to monitorowanie różniły się jednak między gatunkami: szczury opierały się bardziej na dowodach z bieżącej próby, podczas gdy ludzie korzystali ze zgromadzonej historii prób i wyuczonych reguł nagradzania. Co istotne, gdy rozumienie reguł zadania zostało zrównane między gatunkami, różnice w precyzji TEM zmniejszyły się, wskazując, że wspólna zdolność monitorowania jest kształtowana przez dobór strategii i stosowanie reguł, a nie przez fundamentalne różnice w zdolnościach temporalnych.

Zachowanie TEM u różnych gatunków nasunęło kolejne pytanie: jakie informacje leżą u podstaw metapoznawczych ocen czasowania działań motorycznych u ludzi? Sformułowano dwie konkurujące hipotezy. Zgodnie z hipotezą zegara precyzja TEM

maleje wraz ze wzrostem szumu sygnału zegarowego, ponieważ dodatkowa wariancja motoryczna zakłóca odczyt wewnętrznej reprezentacji temporalnej. Zgodnie z hipotezą motoryczną wariancja motoryczna poprawia precyzję TEM, zapewniając punkt odniesienia między zamierzonym a rzeczywistym działaniem motorycznym. W celu rozróżnienia tych hipotez wariancja sygnału motorycznego była eksperymentalnie manipulowana poprzez porównanie warunku naciśnięcia przycisku z warunkiem ruchu joystickiem. Zwiększone zapotrzebowanie motoryczne selektywnie obniżyło precyzję metapoznawczą, nie wpływając na dokładność czasowania motorycznego, co wspiera hipotezę zegara: jednostki oceniają wewnętrzne reprezentacje temporalne, a nie sygnały wyjścia motorycznego. Jest to zgodne z koncepcją metapoznania temporalnego, w której oceny pewności wyższego rzędu odzwierciedlają metapoznawczy odczyt zmienności czasowania, a nie wykonania motorycznego.

Wiedza o tym, że TEM opiera się na wewnętrznych reprezentacjach temporalnych, nie rozstrzyga jednak, czy ocena metapoznawcza jedynie odczytuje sygnały regulujące produkcję interwałów, czy też stanowi neuronie odrębny proces. Trzecie badanie odpowiedziało na to pytanie bezpośrednio, dekodując zoptymalizowaną metodą PCA dane EEG z pojedynczych prób za pomocą modelu Vision Transformer rejestrującego dynamikę fal theta (θ; 3–7 Hz), alfa (α; 8–12 Hz) i beta (β; 13–30 Hz) zarówno podczas fazy czasowania działań motorycznych, jak i fazy decyzyjnej TEM. Dokładność szacowania czasu motorycznego była dekodowalna z dowolnego pojedynczego pasma częstotliwości, zgodnie z niezależnymi kodami oscylacyjnymi specyficznymi dla danego pasma. Wnioskowanie metapoznawcze, zgodność między obiektywną wydajnością czasowania a subiektywną oceną błędu, wymagały natomiast łącznego wkładu wszystkich trzech pasm podczas bieżącego czasowania motorycznego, przy czym żadne pasmo nie było wystarczające samodzielnie. Co więcej, osoby, których aktywność neuronalna przed zakończeniem działania motorycznego bardziej wiarygodnie kodowała stany metapoznawcze, wykazywały większą precyzję TEM. Łącznie wyniki te ustanawiają ocenę metapoznawczą jako neuronie odrębny proces, wspierany przez wielopasmową integrację wyższego rzędu.

We wszystkich trzech badaniach wewnętrzne reprezentacje temporalne stanowiły podstawowy substrat TEM, kształtowany przez reguły zadania, wymagania motoryczne i oczekiwania dotyczące nagrody. Selektywne pogorszenie precyzji TEM przy wyższych

wymaganiach motorycznych wspiera konto oparte na zegarze, w którym ocena błędów odzwierciedla zmienność reprezentacji czasu trwania, a nie wnioskowanie motoryczne. Rozbieżność między specyficznymi dla pasm kodami czasowania motorycznego a wielopasmowym wymogiem wnioskowania metapoznawczego potwierdza, że TEM działa poprzez mechanizmy neuronowe wyższego rzędu, odrębne od monitorowanej przez nie wydajności czasowania.

Słowa kluczowe: metapoznanie, monitorowanie błędu temporalnego (TEM), czasowanie działań motorycznych, produkcja interwałów, dekodowanie EEG, architektura transformera.

Özet

Üstbiliş, bireyin kendi bilişsel süreçlerini izleme ve değerlendirme kapasitesi olarak geniş bir öz-izleme yetileri yelpazesini kapsamaktadır. Zamansal hata izleme (TEM), bu yetilerden biri olarak, bireylerin dışsal geri bildirim olmaksızın kendi ürettikleri zaman aralıklarının doğruluğunu değerlendirdiği üstbilişsel bir süreç olarak tanımlanır. TEM, üstbilişin incelenebilir bir örneği olarak, öz-izlemenin hesaplamalı ve sinirsel altyapısına doğrudan bir pencere sunmaktadır. Motor eylem zamanlaması, TEM'i araştırmak için işlevsel bir deneysel çerçeve sağlar: üretilen zaman aralığı ile hedef arasındaki sapma, içsel olarak erişilebilir bir hata sinyali oluşturur ve bu sayede dışsal geri bildirim olmaksızın üstbilişsel değerlendirmeyi mümkün kılar. Bu tez, türler arası davranışsal paradigmaları, motor talep manipülasyonlarını ve transformer tabanlı EEG mimarisiyle gerçekleştirilen tek deneme düzeyinde sinirsel çözümlemeleri bütünleştirerek TEM'in çok düzeyli bir incelemesini sunmaktadır.

İlk araştırma sorusu, TEM'in türler arasında korunmuş bir kapasite olup olmadığı ve bu kapasitenin paylaşılan mı yoksa farklı stratejilerle mi desteklendiğiydi. Sıçanlar ve insanlar, hizalanmış deneme yapıları, ödül ve görev kuralları ile eşleştirilmiş koşullar altında test edildi; her iki tür de üretilen zaman aralıkları ile hedef arasındaki sapmaya göre doğru ödül seçeneğini belirlemek zorundaydı. Sonuçlar, her iki türün de deneme bazında güvenilir hata izleme sergilediğini göstererek TEM'in paylaşılan bir bilişsel kapasite olduğunu ortaya koydu. Ancak bu izleme sürecini destekleyen bilgi türü türler arasında farklılık gösterdi: sıçanlar ağırlıklı olarak mevcut denemedeki kanıtlara dayalıyken, insanlar birikmiş deneme geçmişi ve öğrenilmiş ödül kurallarından daha fazla yararlandı. Kritik olarak, görev kurallarının türler arasında eşitlendiği durumlarda TEM performansındaki farklar azaldı; bu da TEM'in temel zamanlama yetisindeki farklılıklardan ziyade strateji seçimine ve kural kullanımına bağlı olarak şekillendiğini göstermektedir.

TEM'in türler arasında korunmuş olması, insanlarda motor eylem zamanlamasına ilişkin üstbilişsel yargıların hangi bilgiye dayandığı sorusunu gündeme getirdi. Bu bağlamda iki rakip hipotez formüle edildi. Merkezi Saat Hipotezi'ne göre, saat sinyalinin varyasyonu arttıkça TEM'in kesinliği azalır; çünkü ek motor varyansı içsel zamansal temsillerin okunmasını bozar. Buna karşılık Motor Hipotezi, motor değişkenliğin amaçlanan ve

gerçekleştirilen eylemler arasında bir referans noktası sağlayarak TEM'in kesinliğini artırabileceğini öne sürer. Hipotezleri ayırt etmek için motor sinyal varyansı, düğmeye basma ve joystick hareketi koşulları karşılaştırılarak deneysel olarak manipüle edildi. Bulgular, artan motor talebin motor zamanlama doğruluğunu etkilemeden üstbilişsel kesinliği seçici olarak bozduğunu gösterdi; bu durum, Merkezi Saat Hipotezi'ni destekleyerek katılımcıların değerlendirmelerini motor çıktı sinyalleri yerine içsel zamansal temsillere dayandırdığını ortaya koymaktadır.

TEM'in içsel zamanlama temsillerine dayandığını bilmek, üstbilişsel değerlendirmenin yalnızca motor zamanlama sinyallerine mi dayandığını yoksa nöronal olarak ayrı bir sürece mi sahip olduğunu tek başına açıklamamaktadır. Üçüncü çalışma, bu soruyu doğrudan ele alarak hem motor eylem zamanlamasında hem de TEM karar aşamalarında teta (θ; 3–7 Hz), alfa (α; 8–12 Hz) ve beta (β; 13–30 Hz) bantlarındaki dinamikleri, PCA ile optimize edilmiş tek deneme EEG verilerini Vision Transformer algoritması ile analiz etti. Motor zamanlama doğruluğu herhangi bir tek frekans bandından çözümlenebilirken, üstbilişsel çıkarım, nesnel zamanlama performansı ile öznel hata değerlendirmesi arasındaki uyum, süregelen motor zamanlama sırasında üç bandın ortak katkısını gerektirdi; hiçbir tek bant tek başına yeterli olmadı. Dahası, motor zamanlama süresince sinirsel aktiviteleriyle üstbilişsel çıkarımları daha başarılı bir şekilde çözümlenen bireyler, deneyde daha yüksek TEM karar kesinliği sergiledi. Bu bulgular, üstbilişsel değerlendirmenin motor zamanlama sinyallerinin doğrudan okunmasına değil, üst düzey çok bantlı entegrasyon tarafından desteklendiğine ve nöronal açıdan ayrı bir süreç olarak işlendiğine işaret etmektedir.

Üç çalışma genelinde, içsel zamanlama temsillerinin TEM'in temel substratını oluşturduğu ve görev ile bağlamsal faktörler tarafından şekillendirildiği gösterilmiştir. Artan motor talepler altında TEM kesinliğinin seçici olarak bozulması ve motor zamanlama ile üstbilişsel çıkarım arasındaki frekans temelli ayrışma, TEM'in zamanlama performansından bağımsız, üst düzey sinirsel mekanizmalar aracılığıyla işlendiğini ortaya koymaktadır.

Anahtar kelimeler: üstbiliş, zamansal hata izleme (TEM), motor eylem zamanlaması, zaman aralığı üretimi, EEG kod çözme, transformer mimarisi.

Funding Declaration

This thesis was carried out within the framework of the OPUS project (no. 2019/35/B/HS6/04389), titled "Dynamika procesów neuronalnych u podstaw monitorowania błędów czasowych u człowieka i szczura" ("Dynamics of neuronal processes underlying temporal error monitoring in humans and rats"), funded by the National Science Centre (Narodowe Centrum Nauki, NCN) and awarded to Dr. Tadeusz W. Kononowicz (Institute of Psychology, Polish Academy of Sciences).

Author Contribution Statement

From October 2021 to April 2026, I, Sena Nur Bilgin, worked as a doctoral research assistant on this project. The human studies were conducted at the Institute of Psychology, Polish Academy of Sciences (Warsaw, Poland), while the animal experiments were carried out at NeuroPSI (Institut des Neurosciences Paris-Saclay, CNRS, Université Paris-Saclay, France) in collaboration with Dr. Valérie Doyère. I contributed primarily to the experimental design, data acquisition, preprocessing, analysis, and the preparation of publications across all three studies. Following the conclusion of the project's funding period, I devoted the final two years of my doctoral work to comprehensive data analysis, interpretation of the findings, and the writing of the manuscripts that constitute the core of this thesis.

AI and Software Tools Declaration

Grammar and punctuation were checked using Grammarly. Claude AI (Sonnet 4.6, Anthropic) was used as an assistive tool for identifying citation inconsistencies and structural errors in the manuscript during the revision process. All scientific content, interpretations, and conclusions are solely the work of the author.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Dr. Hab. Tadeusz W. Kononowicz and Dr. Hab. Mateusz Gola, for their continuous support, guidance, and patience throughout my doctoral journey. I am also sincerely thankful to Dr. Valérie Doyère for her exceptional mentorship and for providing me with the opportunity to conduct research internships at the Institute of Neuroscience (NEUROPSI), Paris-Saclay University, supported by NAWA international exchange scholarships.

I am particularly grateful to Dr. Dunia Giomo for her invaluable personal and professional support, as well as her essential contribution to EEG data collection. My thanks also extend to Prof. Dr. Irina Dubrow, Ms. Aleksandra Filipowicz, Dr. Hab. Przemysław Tomalski, and the administrative staff of the Institute of Psychology for their support throughout my studies. I further wish to acknowledge the Scientific Council and the reviewers for their time, effort, and dedication in bringing this doctoral process to completion.

Most importantly, I would like to express my deepest gratitude to my family. Throughout this nearly five-year journey, even across thousands of miles, their unwavering love and encouragement were always within reach. I carried with me the memory of the Bosphorus in Istanbul, the scent of the sea, and the sound of seagulls, as a quiet and enduring reminder of home. For every challenge I have faced along the way, I remain deeply grateful.

Table of Contents

Abstract (EN) …….. ii
Streszczenie (PL) …… iii
Özet (TR) …… iv
Funding Declaration……. v
Author Contribution Statement…… v
AI and Software Tools Declaration……. vi
Acknowledgements ……. vii

1) Introduction 13
1.1) How Does the Brain Monitor Itself? The Role of Motor Action Timing in TEM…… 13
1.2) Cross-Species Insights: Comparative Approach…… 15
1.3) The Clock vs. Motor Debate……..…… 16
1.4) Neural Dynamics of Temporal Metacognition……..…… 17
1.5) Key Contributions: Bridging Behavior and Brain …… 19

2) Methods and Results 20
2.1) Paper 1: Direct Comparison of Temporal Error Monitoring in Humans and Rats……. 20
2.2) Paper 2: Monitoring of Internal Clock or Just Motor Noise? ……... 23
2.3) Paper 3: The Brain as Actor and Evaluator: Distinct Neural Codes for Timing and Self-Evaluation of Timing Errors Revealed by Transformer-Based Decoding 26

3) General Discussion 31
3.1) Limitations and Future Directions……. 34

4) References…….. 36

5) Appendices……. 41
5.1) Paper 1 (Full Publication)
5.2) Paper 2 (Full Publication)
5.3) Paper 3 (Preprint, bioRxiv)

1) INTRODUCTION

1.1) How Does the Brain Monitor Itself? The Role of Motor Action Timing in TEM

How does the brain evaluate the accuracy of its own timing in the absence of external feedback? This question defines temporal error monitoring (TEM), the capacity to assess both the magnitude and direction of self-generated timing errors on a trial-by-trial basis. TEM constitutes a particularly demanding form of metacognition, the broader capacity to monitor and regulate one's own cognitive processes.[1,2] Motor action timing provides a particularly well-suited framework for investigating TEM. Unlike perceptual timing tasks, which rely on externally presented stimuli, motor timing requires participants to generate and sustain temporal intervals internally. This process produces a precise, trial-level error signal (the deviation between intended and actual duration) without any external reference for verification. This property allows metacognitive evaluation to be studied in isolation from externally-driven correction, making error monitoring the only available basis for behavioral adaptation.

The subjective sense of elapsed time plays a central role in structuring perception, guiding the anticipation of future events, and enabling the precise coordination of motor actions,[3] making timing an especially powerful window into how the brain represents and monitors internal state variables. When these monitoring mechanisms are disrupted, the consequences can be substantial. Impairments in evaluating self-generated behavior have been documented in schizophrenia[4] and obsessive-compulsive disorder,[5] where they are associated with poor functional outcomes. Understanding how the brain generates and uses internal error signals is, therefore, both a fundamental scientific question and a clinically relevant one.

How the brain tracks the passage of time has given rise to a range of computational models that differ in the degree to which they represent temporal uncertainty explicitly. Some accounts encode time through intrinsic neural dynamics without representing the reliability of those estimates. State-dependent network models, for instance, propose that duration is encoded in the evolving activity patterns of recurrent neural populations, where each moment corresponds to a unique network state, with no explicit signal for how reliable that state is as a temporal index.[6] Probabilistic models occupy a different position on this spectrum. Scalar Expectancy Theory (SET)[7]

proposes a pacemaker–accumulator mechanism subject to scalar variability. Accordingly, the standard deviation of timing estimates scales proportionally with interval duration. This provides a natural account of Weber's law in the timing domain. Although SET acknowledges this variability, it does not represent uncertainty explicitly as a probability distribution over durations; it quantifies variability as a property of the system rather than encoding it as information the system can consult. Related accumulator frameworks have modeled interval timing as a drift-diffusion process in which a noisy evidence signal is integrated over time until a decision threshold is reached, with timing precision determined by the rate and variability of that accumulation.[8] Importantly, while drift-diffusion accounts can recover a form of confidence from the dynamics of the accumulator itself, this confidence signal is not independent of the timing process; it is an emergent property of the same first-order mechanism rather than a separately represented quantity available for higher-order evaluation. Bayesian models go further by treating temporal uncertainty as a quantity the brain actively represents and exploits as a distinct computational variable. They propose that the brain integrates a noisy duration measurement, whose uncertainty grows with interval length in accordance with scalar variability, with a learned prior over the temporal context to derive optimal estimates.[9,10] A key behavioral signature of this integration is the central tendency effect: reproduced intervals regress systematically toward the mean of the experienced distribution, and this bias grows stronger for longer intervals, precisely because longer measurements are less reliable and the prior therefore exerts greater influence on the posterior.[9]

Whether a timing model represents uncertainty explicitly has direct consequences for TEM. Without such a representation, a system that only tracks elapsed time lacks any internal basis for evaluating accuracy. In contrast, a system that also encodes the reliability of its estimates carries precisely the information required for metacognitive evaluation. Critically, this uncertainty signal is not random noise. Instead, it scales lawfully with interval duration and task demands. This regular structure makes it a well-defined quantity that could, in principle, be read out and used to guide error monitoring. Consistent evidence from direct error monitoring paradigms indicates that both humans[11-14] and rodents[15] access this signal on a trial-by-trial basis, and adjust reward expectations according to the estimated precision of their temporal productions. Complementary behavioral work further suggests that this sensitivity to temporal uncertainty also shapes implicit decision-making in probabilistic timing contexts across

species[16], and that organisms can optimally calibrate their responses under temporal risk by integrating uncertainty about interval duration with expected reward structure.[17]

Yet despite these convergent lines of evidence, no existing timing model explicitly incorporates a dedicated mechanism for TEM as a distinct process separable from the timing mechanism itself. This gap defines the scope of the present thesis and motivates its three fundamental questions. The first concerns whether TEM is a conserved capacity, specifically, whether trial-by-trial evaluation of self-generated timing errors is uniquely human or shared with non-human animals in a translational behavioral paradigm. The second concerns the informational basis of TEM: whether error monitoring draws on internal clock representations, or whether it instead derives from peripheral motor feedback, such as proprioceptive or efference copy signals generated during movement execution. The third, and perhaps most fundamental, question concerns the neural organization of TEM: whether error monitoring arises from the same decision variable as interval production, or whether it constitutes a higher-order process that is at least partially distinct from timing performance itself. Together, these three questions define a research framework aimed at understanding not merely how the brain represents and produces temporal intervals, but how it comes to know, and act upon, the reliability of those representations.

1.2) Cross-Species Insights: Comparative Approach

Understanding temporal error monitoring (TEM) from a comparative perspective is essential for determining whether this capacity is uniquely human or reflects a more broadly conserved cognitive function. Examining how different species monitor the accuracy of their own timing behavior allows researchers to identify shared behavioral signatures that may reflect domain-general performance monitoring mechanisms. Early rodent studies demonstrated that rats express uncertainty in duration discrimination tasks by opting for "safe" choices when stimuli fall near the discrimination threshold.[18] More recent work has shown that rats can report timing errors on a trial-by-trial basis and adjust reward expectations according to their internal assessments of accuracy.[15] Together, these findings suggest that TEM may reflect an evolutionarily conserved capacity for performance evaluation rather than a uniquely human ability.[15,17] However, low-level associative accounts have also been proposed to explain similar uncertainty-monitoring

behaviors in rodents,[19] meaning that convergent behavioral signatures do not necessarily imply shared underlying mechanisms.

Drawing cross-species inferences requires careful consideration of both construct validity and computational validity. Construct validity concerns whether a theoretical construct is defined and operationalized equivalently across species, for instance, by aligning error-monitoring behaviours through matched trial structures and consistent decision rules. Computational validity, in contrast, concerns how task-relevant rules are internally represented and implemented, acknowledging that similar observable behaviours may arise from distinct underlying algorithmic processes across species. These two dimensions must be considered jointly, as even subtle variations in experimental design, or differences in how task contingencies are learned and represented, can lead to divergent interpretations.

Building on this framework, the first paper[20] of this thesis investigated trial-by-trial TEM in both species using a closely matched paradigm adapted from rats[15] to humans. Rather than targeting shared neural substrates or specific mechanistic processes, the aim was to characterize behavioral profiles across species and extract rule-based variables that could reveal convergent patterns in task performance. Specifically, we examined whether error monitoring could be reliably expressed and quantified in both species, and how motor action timing metrics, such as temporal precision and accuracy, related to error judgments and response times across species.

1.3) The Clock vs. Motor Debate

A core theoretical question in TEM concerns the informational source on which error monitoring draws. The Wing–Kristofferson decomposition[21] proposes that timing variance can be formally separated into a central clock component and a peripheral motor component. Applied to performance monitoring, this separation generates two competing hypotheses. The Clock Hypothesis holds that TEM draws on the central timing signal. Under this account, increased motor demands impair metacognitive precision because additional motor variance is not part of the internal duration representation. The Motor Hypothesis, by contrast, holds that TEM draws on motor feedback or efference copies of motor actions. Under this account, increased motor demands make errors more salient and could even enhance metacognitive discrimination. Rather than treating these as mutually exclusive, we consider them as opposite ends of a continuum along which

individuals may draw on both sources of information. Empirically, these contributions can be partially dissociated: if metacognitive precision is selectively impaired by increased motor demands without a corresponding change in timing accuracy, this would suggest that clock-based information plays a relatively greater role than peripheral motor feedback in grounding TEM judgments.

Empirically addressing this question requires manipulating motor signals while measuring timing errors and confidence across tasks with varying sensorimotor demands. The second paper[22] of this thesis was designed with this goal. Participants performed both time production and reaction time tasks, with motor demands varied through button-press and joystick-movement conditions across dominant and non-dominant hands. The study pursued two complementary aims: first, to determine whether TEM draws primarily on internal clock signals or on motor execution feedback, by testing how increased motor signal demands affect error monitoring and confidence; and second, to assess whether introspective access to timing errors generalizes across distinct timing paradigms.

1.4) Neural Dynamics of Temporal Metacognition

Even if behavioral evidence favors the Clock Hypothesis, this does not fully resolve the role of the motor system in TEM. Ruling out peripheral motor feedback does not rule out that neural dynamics conventionally classified as motor contribute to metacognitive evaluation at a higher level. Beta-band oscillations illustrate this ambiguity. β activity is conventionally associated with motor execution and post-movement rebound, yet it also encodes temporal intervals at the single-trial level[23,24] and has been implicated in sensorimotor state maintenance,[25] content-specific memory reactivation,[26] and confidence in internal motor model estimates.[27] β, therefore, cannot be straightforwardly assigned to either motor execution or metacognitive evaluation. The central question is whether, when the brain evaluates a timing error, it recruits the same oscillatory signals that produced the timed interval or engages a distinct set of neural representations.[28,29]

Prior studies have provided partial evidence bearing on this debate. Beta-band (β; 13–30 Hz) oscillatory dynamics have been implicated as correlates of internally generated timing, with the separability of single-trial neural trajectories predicting both temporal accuracy[23,24] and metacognitive sensitivity.[29] Alpha-band (α; 8–12 Hz) activity following interval offset has been shown to track participants' signed self-evaluations in

a graded, linear fashion, consistent with a post-interval inferential readout rather than a simple motor error signal.[29] Frontal midline theta has been implicated in temporal prediction errors, with greater phase coherence observed as error magnitude increases.[30] Yet whether these frequency-specific mechanisms are shared between motor action timing performance and metacognitive evaluation, or whether the latter engages neurally independent representations, cannot be resolved from condition-averaged analyses alone.

Resolving this requires a theoretical framework that specifies what relationship between performance and metacognitive evaluation is expected, and what dissociation patterns would constitute evidence against each position. Theories of metacognition differ in how they conceptualize this relationship. Single-process accounts propose that confidence and performance emerge from a common decision variable, with metacognitive judgments arising directly from the same decision variable that guides task performance, without invoking additional computational mechanisms.[31] Whether this prediction holds empirically can be assessed using signal detection theory, originally formalized as a framework for separating perceptual sensitivity from response bias,[32] and since extended to quantify metacognitive sensitivity relative to first-order performance.[33,34] These tools allow researchers to assess whether confidence-accuracy relationships depart from a shared-signal baseline, that is, whether metacognitive judgments efficiently reflect first-order performance signals or fall systematically short of what performance alone would predict. Where such a shortfall is observed, it suggests that metacognition does not arise simply from the same decision variable as performance, a possibility that motivates accounts in which metacognition constitutes an at least partially distinct, albeit dependent, processing stage.

Hierarchical accounts occupy precisely this middle ground: metacognition is treated as a distinct processing stage that nonetheless operates on first-order performance signals, such that stronger neural encoding of performance produces more precise metacognitive judgment.[29,35] However, hierarchical accounts still tie metacognitive precision to the quality of first-order performance encoding, leaving open the question of whether metacognitive evaluation can operate through representational processes that are genuinely independent of, rather than merely downstream from, first-order performance signals.

Higher-Order Representation theory (HOR) addresses this question most directly, arguing that conscious awareness of a mental state requires a higher-order representation that explicitly takes that state as its object, rather than simply arising from it.[36] Applied

to metacognition, this implies that error evaluation must involve a distinct representational level, a claim supported neurally by evidence that dedicated prefrontal circuits contribute to higher-order representations in ways that cannot be reduced to first-order processing circuits.[28,37]

These theoretical positions generate concrete, competing predictions about the neural organization of TEM that require single-trial resolution to test. Condition-averaged analyses, by collapsing neural activity across trials, cannot capture whether timing performance and metacognitive evaluation share or dissociate their neural substrates on a trial-by-trial basis. The third paper[38] of this thesis addressed this directly by decoding single-trial EEG data recorded during both motor action timing and subsequent error evaluation, using a Vision Transformer-based architecture.[39,40] Three related questions guided the analysis: whether single-trial neural dynamics predict timing accuracy; whether the neural representations underlying metacognitive evaluation dissociate from those supporting timing performance; and whether any such dissociation is detectable across multiple frequency bands spanning both the timing and TEM phases of the task.

1.5) Key Contributions: Bridging Behavior and Brain

Each of the three papers comprising this thesis addresses a distinct question about TEM, together forming a progression from behavioral characterization to informational dissociation to neural organization. The first paper[20] established TEM as a conserved capacity by demonstrating reliable trial-by-trial error monitoring in both rats and humans under matched experimental conditions. Notably, cross-species differences in performance were attributable to differences in strategy use and rule representation rather than to any fundamental disparity in timing ability. Building on this, the second paper[22] dissociated the informational basis of TEM by showing that increased motor signal demands selectively impaired metacognitive precision without affecting timing accuracy, supporting a clock-based rather than motor-based account of TEM. The third paper[38] then addressed the neural organization of this capacity, demonstrating that metacognitive evaluation does not arise directly from timing performance signals alone but instead requires coordinated multi-band neural integration, establishing TEM as an at least partially distinct higher-order process. Together, these findings advance our

understanding of how the brain not only represents and produces temporal intervals, but also comes to evaluate the reliability of those representations.

2) METHODS AND RESULTS

The following section presents focused summaries of the methods and principal findings of each study, organized to reflect the sequential logic of the thesis. Procedural details, statistical formulations, and main and supplementary analyses are reported in full in the original publications; what is presented here is limited to the aspects most directly relevant to the theoretical questions motivating each study and to the interpretive continuity across the three papers.

2.1) Paper 1: Direct Comparison of Temporal Error Monitoring in Humans and Rats

Subjects/Apparatus: The human sample comprised 34 participants (18 women, 16 men; *M*age = 26.89 years, *SD* = 6.50 years) with normal or corrected-to-normal vision and hearing. All participants provided written informed consent prior to testing and received compensation in the form of points convertible to virtual gift cards. The animal sample comprised 16 male Sprague–Dawley rats aged three months at the start of testing. Rats were group-housed under a 12-hour light/dark cycle, food-restricted to 85% of free-fed body weight, and handled daily throughout the study. Rats had prior experience with a 3.2-second timing task[15] and were retrained on a 2-second target interval to match the duration used in the human task. Both species completed two daily sessions separated by three-hour breaks. All procedures received ethical approval from the relevant institutional review boards for both human and animal research. Human participants completed the task on a desktop computer using MATLAB (MathWorks) with the Psychophysics Toolbox. Rats were tested in standard operant chambers, with lever presses, port entries, and reward delivery recorded with millisecond precision.

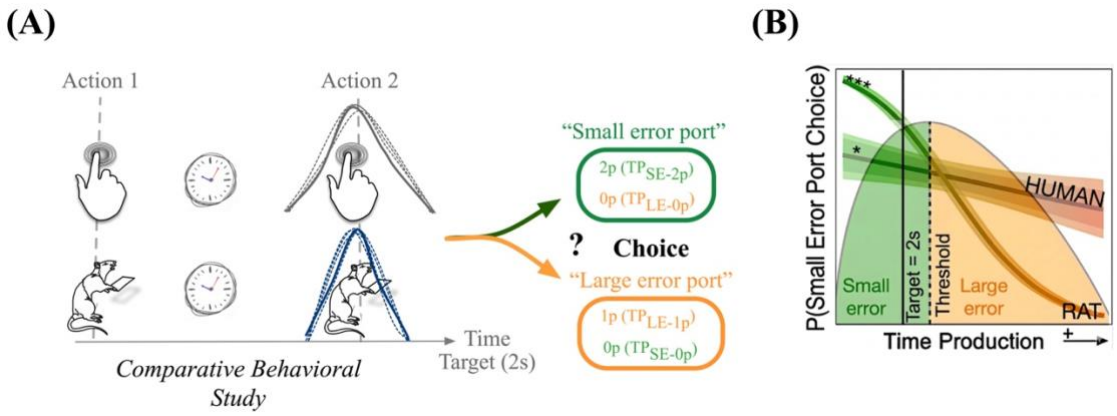


Figure 1. Cross-Species Temporal Error Monitoring Paradigm and Behavioral Performance

(A) Task structure used in the comparative behavioral study. Both humans and rats produced a 2-second target interval (Action 1 to Action 2) and subsequently chose between a "Small error port" (high reward: 2p on accurate trials, 0p on large-error trials) and a "Large error port" (low reward: 1p on large-error trials, 0p on accurate trials). Correct port selection reflected successful trial-by-trial temporal error monitoring.
(B) Choice Probability as a Function of Time Production. The figure depicts the probability of selecting the small error port as a function of time production (TP) for humans and rats, with shaded regions indicating 95% confidence intervals. The solid vertical line marks the 2-second target, and the dashed vertical line marks the reward rule threshold separating small from large errors. Rats exhibited a steeper response function, reflecting greater sensitivity to current-trial TP error, whereas humans showed a shallower, more distributed curve, consistent with greater reliance on accumulated task-rule representations. Nevertheless, both species showed a systematically decreasing probability of small error port choice as their TPs deviated further from the target, demonstrating reliable trial-by-trial TEM across species.

Procedure: The TEM task was adapted from an established rat protocol[15] to enable direct cross-species comparison. Both species were trained to produce a target interval and evaluate their time production (TP) errors across three successive phases: Duration Production Training, TEM Training, and TEM Test (Figure 1A). Rats first underwent a retraining phase in which the target interval was reduced from 3.2 to 2 seconds over four Duration Production Training sessions, followed by 23 TEM Training sessions and 34 TEM Test sessions of 120 trials each. Human participants completed two daily sessions

of 100 trials each, beginning with pre-training trials to familiarize them with the target interval. Participants were instructed to avoid external timing strategies throughout.

The human TEM training phase comprised two instruction conditions: an Instruction condition, in which participants received detailed explanations of the reward rules, and a No-Instruction condition, in which guidance was kept minimal. These were crossed with two interval production methods (HOLD and PRESS), resulting in four experimental groups. During TEM Training, a probabilistic reward structure linked error magnitude to reward outcome: small TP errors were associated with higher rewards and large errors with lower rewards. To reinforce this association, only 50% of trials delivered the expected reward from the corresponding port; on the remaining trials, no reward was delivered from the opposite port, encouraging rats and humans to learn the relationship between TP accuracy, reward magnitude, and port assignment (Figure 1A).

During the TEM Test trials, rats and humans selected the card or port corresponding to their TP accuracy on 30% of trials, with correct selections indicating successful TEM. Human participants could earn up to 130 points per session; rats could earn up to 144 food pellets per session.

Data Analysis: Continuous variables, TP accuracy (μ), precision (σ), and response times (RTs), were analyzed using Linear Mixed-Effects Models (LMMs). Binary outcomes (port/card choice and choice accuracy) were analyzed using Generalized Linear Mixed-Effects Models (GLMMs). All models included random effects for participants and sessions, with fixed effects for experimental conditions. Model selection was guided by the Akaike Information Criterion (AIC).

To enable cross-species comparisons, human sessions were aligned with rat sessions using theta (θ) alignment indices, which quantified how closely each subject's internal representation of task rules matched the underlying reward structure. Three indices were defined: Assigned θ, reflecting the threshold separating small from large errors based on each subject's prior TP distribution; Observed θ**,** reflecting the same threshold based on the current session's TP distribution; and Choice θ, capturing the decision boundary inferred from port selections (small vs. large error) within the current session. These indices were incorporated into GLMMs as trial- and session-level predictors of choice accuracy.

Results: Both species successfully produced the target intervals, with rats demonstrating higher accuracy and precision than humans. In both species, TP variability increased with interval length, consistent with Weber's law. Both species monitored TP errors and adjusted their choices accordingly: the probability of selecting the small error port decreased with longer current-trial productions. Rats relied more heavily on current-trial information and exhibited higher overall choice accuracy (Figure 1B).

Session-level theta alignment strongly predicted choice accuracy in both species, and cross-species differences in choice accuracy were substantially reduced after matching for task-rule representation. At the trial level, rats monitored productions relative to all three theta distances, whereas humans relied primarily on the Assigned θ, reflecting greater dependence on the learned reward rule. Both species responded faster when anticipating higher rewards. However, rats' RTs were more closely linked to choice accuracy than those of humans, suggesting that RTs served as a more reliable index of trial-by-trial metacognitive confidence in rats.

2.2) Paper 2: Monitoring of Internal Clock or Just Motor Noise?

Subjects/Apparatus: 31 human participants (19 women, 12 men; *M*age = 25.70 years, *SD* = 5.74 years) with normal or corrected-to-normal vision and hearing participated in the study. The protocol was approved by the Ethics Committee of the Institute of Psychology, Polish Academy of Sciences, and participants received 100 PLN compensation. All participants provided written informed consent before the experiment. Each session lasted approximately 1.5 hours and comprised one pretraining phase, two training blocks, and three experimental blocks (two TP and one RT). Participants responded using a joystick, performing grasping and pulling movements in MOVE trials and button presses in PRESS trials (Figure 2A). Stimuli were presented via MATLAB with the Psychophysics Toolbox, and behavioral data were exported for analysis in R.

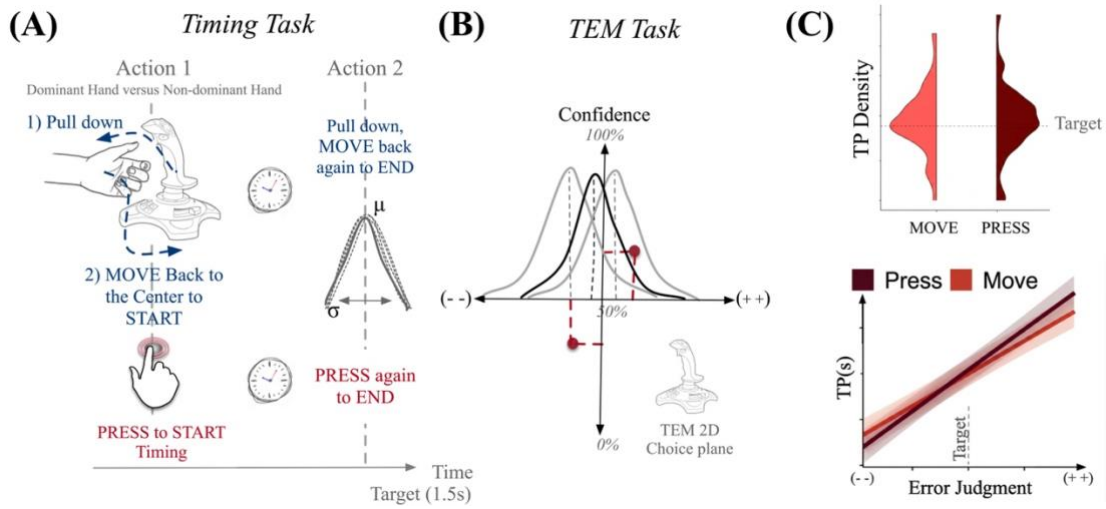


Figure 2. Motor Demand Manipulation Paradigm and Behavioral Performance Across PRESS and MOVE Conditions

(A) Timing task structure illustrating the two motor demand conditions. In the MOVE condition (blue), participants initiated timing by pulling the joystick downward and returning it to the center (Action 1), then repeated the movement to terminate the interval (Action 2). In the PRESS condition (red), participants initiated and terminated the interval via button presses. Both conditions targeted a 1.5-second interval, with dominant and non-dominant hand use counterbalanced across participants. Distributions at Action 2 reflect the trial-by-trial variability in time production, where μ denotes the mean and σ the standard deviation.

(B) TEM two-dimensional response interface. Following each time production trial, participants reported both the direction and magnitude of their timing error along the horizontal axis (ranging from *too short* (− −) to *too long* (+ +)) and their confidence in that judgment along the vertical axis (0–100%). The red dashed lines depict two example responses: one reflecting low confidence paired with a slightly early error judgment, and one reflecting high confidence paired with a slightly late error judgment.

(C) Behavioral results across motor demand conditions. The upper panel shows TP density distributions for MOVE (light red) and PRESS (dark red) conditions relative to the 1.5-second target (dashed line), indicating comparable timing accuracy across conditions. Lower panel shows the correlation between error judgment and TP duration for PRESS (dark red) and MOVE (light red) conditions, with shaded regions indicating 95% confidence intervals. The steeper slope in the PRESS condition reflects greater

metacognitive precision, indicating that lower motor demand supported more precise TEM.

Procedure: The study employed a 2×2 within-subjects design, manipulating motor demand (PRESS vs. MOVE) and responding hand (dominant vs. non-dominant). During pretraining, participants familiarized themselves with the 1.5-second target interval and were instructed to avoid explicit counting strategies throughout. In the TP task, PRESS trials required two joystick button presses to mark interval onset and offset, whereas MOVE trials required a downward joystick movement followed by a return to center to indicate the same temporal boundaries. The responding hand was indicated by on-screen spatial cues. TP durations outside the 0.5–3.0 s range were considered invalid and excluded from analysis. In the RT task, participants responded to a visual stimulus using either a button press (PRESS) or a downward joystick movement toward the stimulus (MOVE) (Figure 2A).

During training, trial-by-trial feedback was provided, and trials with extreme deviations were restarted. Feedback was expressed as a signed rating reflecting the magnitude and direction of the timing error relative to the participant's own TP distribution. Productions falling 2–3 *SD* from the mean were rated ±0, indicating productions that were *too short* or *too long*. Those falling 1–2 *SD* were rated ±1, 0.5–1 *SD* were rated ±2, and productions within 0.5 *SD* were rated ±3, reflecting increasingly accurate TP.

During the test phase, participants reported both the magnitude and direction of their timing error, along with their confidence in this judgment, using a two-dimensional response interface. The x-axis represented signed error, while the y-axis indexed confidence in this TEM judgment on a scale from 0 to 100% (Figure 2B). Each participant completed two TP blocks and one RT block, with task order counterbalanced across participants.

Data Analysis: Behavioral measures included TP durations, decision times (DTs), error judgment (EJ), and confidence ratings. Confidence ratings were scaled from 0 to 100% and categorized into High and Low groups via median split. All measures were analyzed using Linear Mixed-Effects Models (LMMs) with participants as random effects and experimental conditions as fixed effects. Model selection was guided by AIC.

Results: TP accuracy and precision were comparable across PRESS and MOVE conditions, confirming that motor demand did not affect timing performance (Figure 2C). In the RT task, responses were faster in PRESS than MOVE trials, consistent with lower motor signal demands. EJ ratings were closer to the target in PRESS trials and showed stronger correlations with TP performance (Figure 2C), indicating higher TEM precision when motor signal demands were lower. This pattern was consistent with the Clock Hypothesis: increased motor signal added variance that selectively degraded TEM precision without affecting timing accuracy.

In addition, confidence tracked metacognitive performance closely; high-confidence trials showed stronger EJ–performance correlations across both tasks, suggesting that confidence reflected genuine metacognitive sensitivity. PRESS trials and dominant-hand responses were associated with higher confidence and shorter decision times, indicating faster and more certain metacognitive judgments under lower motor demand conditions. Together, these findings indicate that motor signal demands selectively constrain TEM, and that confidence provides a reliable behavioral index of metacognitive precision.

2.3) Paper 3: The Brain as Actor and Evaluator: Distinct Neural Codes for Timing and Self-Evaluation of Timing Errors Revealed by Transformer-Based Decoding

Subjects/Apparatus: A total of 29 participants (17 male, 12 female; M = 24.10, SD = 4.28) took part in the study. All participants provided written informed consent prior to testing. To ensure neural data quality uncontaminated by stimulation effects, analyses were restricted to the post-Sham EEG-only block, representing approximately 25% of the full dataset, as tDCS can produce aftereffects that persist beyond the stimulation period. The full experiment included a tDCS session and a Sham session, conducted on separate days to reduce cognitive fatigue.

Each session comprised two blocks of approximately 20 minutes, separated by a 5-minute break, with the first block consisting of either active tDCS or Sham stimulation counterbalanced across participants, and the second block consisting of an EEG-only recording. During tDCS, continuous stimulation was delivered at 1 mA; during Sham, current was applied briefly at onset and then discontinued to simulate the stimulation experience. Event markers were transmitted from MATLAB to the EEG system to

annotate trial-critical time points, enabling the extraction of pre-event, event, and post-event epochs relative to the TP and TEM phases.

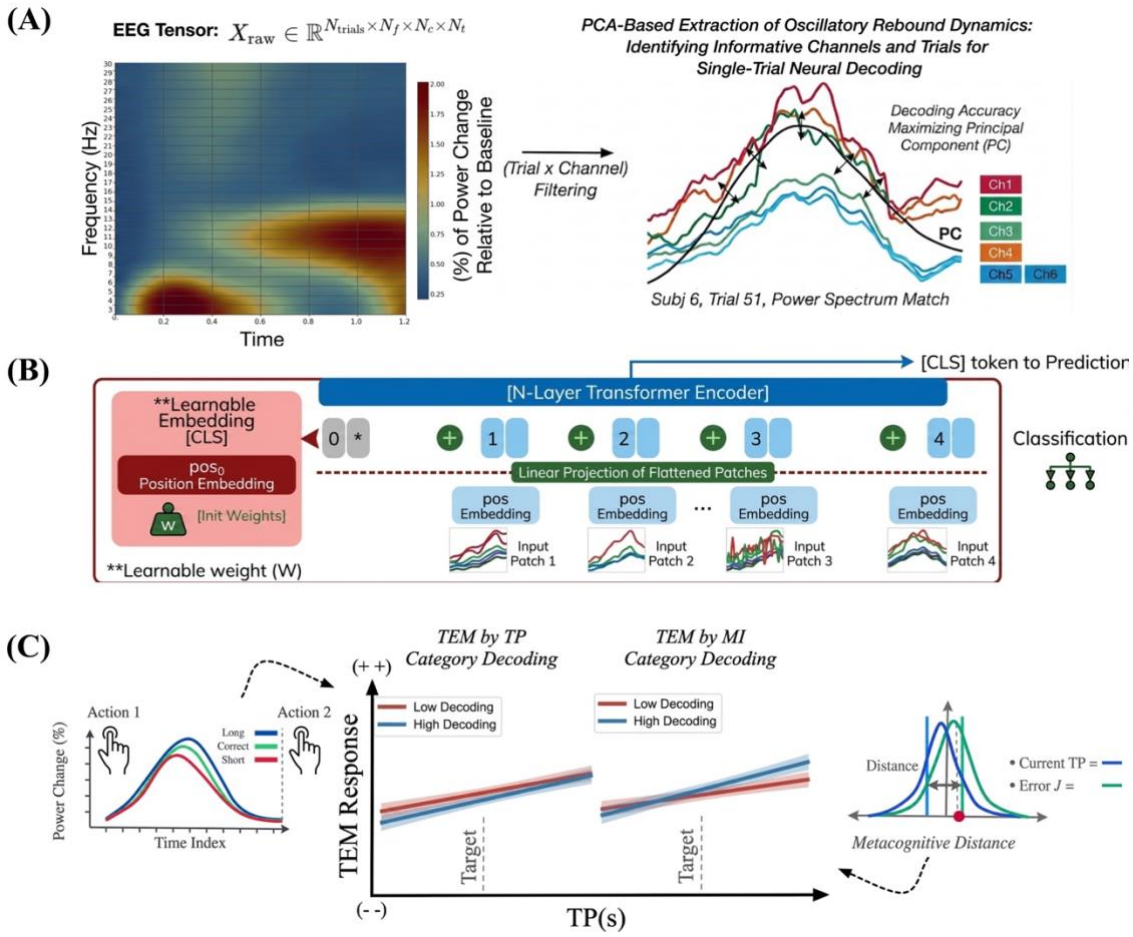


Figure 3. EEG-ViT Pipeline: PCA-Based Trial and Channel Selection, Transformer Architecture, and Single-Trial Decoding of Motor Timing and Metacognitive Inference

(A) PCA-based extraction of oscillatory rebound dynamics for informative channel and trial selection. The right panel displays a time-frequency representation averaged across participants and trials, with color indicating percentage power change relative to a pre-onset baseline (−0.3 to −0.1 s before TP onset). PCA was applied to the raw EEG 4D tensor following trial-by-channel filtering to extract dominant patterns of temporal variance. The component yielding the highest decoding accuracy, typically the first or second, depending on the participant, was selected as the reference component. Channels and trials were then ranked according to their alignment with this component, and the most informative subset was retained for downstream decoding. The black curve

illustrates the selected component's temporal profile across representative channels (Ch1–Ch6), shown here for Subject 6, Trial 51.

(B) Vision Transformer (EEG-ViT) architecture. PCA-filtered single-trial EEG data were segmented into four non-overlapping temporal patches, each spanning approximately 300 ms of the 1.2-second interval, and projected into patch embeddings via linear projection, with learned positional embeddings added at each token position. A learnable classification token, prepended to the patch sequence at position 0 with its own positional embedding, was passed through an N-layer transformer encoder with residual connections. The resulting token representation was forwarded to a multilayer perceptron (MLP) head for single-trial classification of TP, MI, and TEM confidence.

(C) Single-trial decoding results and neural–behavioral dissociation. The left panel shows frequency band-specific oscillatory power change (%) across three time production categories (Long, Correct, Short), illustrating separable neural trajectories preceding motor action termination: longer interval categories are associated with progressively stronger oscillatory activation, consistent with previously established findings. The middle panels present the critical dissociation. Participants were stratified into high and low groups based on the decoding accuracy of their pre-termination neural activity during ongoing motor action timing, separately for timing performance and metacognitive inference. Stratifying by TP decoding accuracy yielded no between-group difference in TEM precision, whereas stratifying by MI decoding accuracy revealed a significant difference, with the high MI decoding group showing substantially stronger correspondence between objective TP and subjective error judgment. This selective pattern demonstrates that neural encoding of metacognitive states during ongoing motor action timing predicts TEM precision independently of timing performance itself. The right panel illustrates the Metacognitive Distance metric, defined as the absolute difference between the current time production (blue) and the subjective error response (green), where smaller distances reflect greater metacognitive inference precision.

Procedure: Training proceeded in two stages. In the first stage, participants learned the 2-second target interval over 10 visual familiarization trials, with an additional 10 trials administered if reliable memorization had not been achieved. They then completed 30 TP trials reproducing the interval with mouse clicks, with trial-by-trial feedback indicating whether each production fell within ±1000 ms of the target. Productions exceeding this threshold were excluded from analyses.

In the second stage, TEM training comprised 10 trials in which participants made two successive metacognitive judgments on each trial: a temporal error response using a circular arc interface, followed by a confidence rating on a 0–100% scale, both consistent with the procedure used in Paper 2, except the response scale design. Personalized deviation feedback was provided on every trial, calibrated to each participant's own timing variability and categorized into four levels: *very close to the target* for deviations within 0.3 *SD*; *a little bit short/long* for deviations between 0.3 and 1 *SD*; *quite short/long* for deviations between 1 and 2 *SD*; and *too short/long* for deviations exceeding 2 *SD*. On Day 2, an additional 15 TEM training trials were administered with feedback recalibrated using the TP standard deviation from Day 1.

In the main experimental phase, each trial comprised a TP task followed immediately by a temporal error response and confidence rating, separated by a 300 ms intertrial interval. Feedback was present on 50% of trials and absent on the remaining 50%, constituting a within-session manipulation of feedback availability. On feedback-present trials, the same personalized deviation categories used during TEM training were applied. On feedback-absent trials, participants relied exclusively on their internal timing representations to evaluate their own performance, providing a direct measure of TEM in the absence of externally driven correction.

Data Analysis: EEG data were preprocessed individually per participant, including artifact rejection, independent component analysis (ICA) cleaning, band-pass filtering (1–120 Hz), 50 Hz notch filtering, and baseline normalization. Preprocessed data were organized as 4D tensors (trials × channels × frequencies × time) and frequency-averaged to reduce dimensionality while preserving temporal dynamics. Principal component analysis (PCA) was applied to identify dominant temporal patterns; the top component most predictive of behavioral outcomes (accuracy-maximizing component) was used to rank channels and trials (Figure 3A). Downstream modeling used the top 10% of channels across three trial retention thresholds (30%, 50%, and 100%).

For single-trial decoding, a Vision Transformer[39](ViT) architecture was employed. Each trial was segmented into non-overlapping temporal patches, yielding compact embeddings of local oscillatory activity. Multi-head self-attention captured relationships across patches, while a classification token ([CLS]) and positional encodings summarized trial-level representations. Embeddings were processed through 6 transformer encoder layers with residual connections, layer normalization, and

feedforward networks, and the final [CLS] token was passed to a multilayer perceptron (MLP) classifier (Figure 3B).

Behavioral metrics included TP, temporal error response, TEM confidence, and Metacognitive Distance (MD). MD was defined as the absolute difference between z-scored temporal error responses and TP durations, and served as an index of Metacognitive Inference (MI). Smaller MD values indicate closer correspondence between objective performance and subjective error reports (i.e., higher MI), whereas larger values reflect greater divergence (i.e., lower MI; Figure 3C, right panel).

Continuous TP values were discretized into three categories, short, correct, and long, based on the 33rd and 66th percentile thresholds (Figure 3C, left panel). MD and TEM confidence were each dichotomized using a median split, resulting in balanced binary classes for MI (high vs. low) and confidence (high vs. low).

EEG data recorded during the TP phase were used to predict TP category, MI, and TEM confidence. In addition, MI was also predicted using EEG recorded during the TEM phase. All models were trained within-subject using stratified 5-fold cross-validation to ensure balanced class representation and to prevent data leakage, with performance averaged across folds.

Results: Single-trial TP categories were successfully decoded from pre-termination EEG recorded immediately before motor action offset using the EEG-ViT architecture. Decoding performance peaked when retaining the top 30% of PCA-ranked trials, exceeding chance levels and outperforming larger trial subsets. The β and θ bands contributed most to TP decoding, while multi-band integration did not improve performance, indicating that single-frequency bands are sufficient for motor timing classification.

In contrast, MI was decodable both during ongoing motor timing and during the TEM decision phase, with stronger predictive performance observed in the latter. During motor execution, reliable MI classification required the integration of θ, α, and β signals, as no single band alone was sufficient. This dissociation, single-band sufficiency for TP versus multi-band dependence for MI, suggests that motor timing and metacognitive evaluation rely on partially distinct neural architectures. Consistently, in the TEM phase, the highest MI decoding accuracy again depended on the joint contribution of θ, α, and β bands.

A neural–behavioral dissociation further challenged single-process accounts. Participants stratified by MI decoding accuracy showed no differences in timing accuracy or variability, yet the correspondence between objective TP performance and subjective error reports was significantly stronger in the high MI decoding group. No comparable effect was observed when stratifying by TP decoding accuracy (Figure 3C, middle panels). This pattern contradicts single-process models, which predict shared neural substrates for performance and confidence, and instead supports the view that metacognitive evaluation constitutes an at least partially distinct representational layer that draws on first-order timing signals while engaging additional evaluative mechanisms.

3) GENERAL DISCUSSION

Across three studies operating at complementary levels of analysis, this thesis investigated the behavioral and neural mechanisms by which humans and non-human animals monitor temporal errors and form metacognitive judgments about motor action timing. The findings established TEM as a conserved cognitive capacity: both rats and humans monitored timing errors trial by trial, yet the strategies supporting this shared capacity diverged systematically, with rats drawing primarily on current-trial evidence and humans relying on accumulated rule representations. This divergence was substantially reduced when task-rule alignment was equated across species,[20] indicating that strategy selection rather than timing ability was the primary determinant of cross-species performance differences, consistent with theoretical accounts that separate the capacity for error detection from its higher-order evaluation.[12] What the cross-species findings reveal, at a broader level, is that the capacity to monitor self-generated temporal errors is not uniquely tied to the cognitive architecture of a species; it is instead a functional property that can be instantiated through different behavioral strategies depending on the representational resources available to the organism.

Having established that TEM is a conserved capacity, the question of what information grounds metacognitive judgments about motor action timing in humans became central. The Wing-Kristofferson model[21] proposes that timing variance can be formally separated into a central clock component and a peripheral motor component, generating two competing predictions: the Clock Hypothesis, under which metacognitive precision depends on the central timing signal, and the Motor Hypothesis, under which

it depends mostly on peripheral motor feedback. A selective dissociation was observed: increasing motor signal demands impaired TEM precision without affecting timing accuracy,[22] a pattern more consistent with the Clock Hypothesis and difficult to accommodate under SET,[7] where if motor noise degraded the timing signal, timing accuracy should have suffered too. A Bayesian framework[9,10] offers one possible interpretation. Increased motor variance may reduce the metacognitive system's ability to distinguish uncertainty arising from the internal clock from variability introduced by motor execution. As a result, access to clock uncertainty signals is selectively degraded, while the interval estimate itself, reflecting the central tendency of productions, remains largely intact. This account is also consistent with the meta-d′ framework[33,34], where a transmission loss between a performance signal and its metacognitive readout indicates at least partially separable processes. Notably, this dissociation was specific to peripheral motor feedback and left open whether motor-related neural dynamics contribute to TEM at a higher level, consistent with evidence that β oscillations index confidence in internal model estimates[27] and encode temporal intervals at the single-trial level.[23,24]

The neural findings of Paper 3 extended the earlier conclusions in three directions, each bearing on a distinct theoretical question. First, metacognitive inference was decodable from pre-termination EEG, prior to any explicit evaluation, indicating that information relevant to error assessment is already encoded in neural activity during interval production rather than arising solely at the stage of metacognitive judgment.[38] This is consistent with a Bayesian account,[9,10] under which uncertainty about the current production accumulates continuously during interval generation and is available as a structured signal prior to any explicit evaluative response. Second, the band-specificity of decoding revealed a qualitative dissociation between timing performance and metacognitive inference: while timing accuracy was decodable from any single frequency band, reflecting independent frequency-specific oscillatory codes, metacognitive inference required the simultaneous integration of θ, α, and β.[38] This finding is not merely a quantitative difference in decoding strength; it indicates that the neural representations supporting metacognitive evaluation are organized differently from those supporting timing performance, drawing on coordinated multi-band dynamics rather than any single oscillatory substrate. The contribution of β to this integration is particularly noteworthy: β participated in metacognitive inference only within the multi-band context, not as an independent motor readout,[38] consistent with proposals that β reflects the maintenance and reactivation of internal representations beyond simple

execution feedback,[25,26] and with direct evidence that post-movement β indexes confidence in internal model estimates in sensorimotor contexts.[27] Third, and most critically for theoretical adjudication, a selective neural–behavioral dissociation was observed: participants whose pre-termination neural activity more reliably decoded metacognitive states showed stronger correspondence between objective timing performance and subjective error evaluation, whereas participants stratified by TP decoding accuracy showed no equivalent difference in either TEM precision or MD.[38] This pattern directly challenges single-process accounts,[8,31] which predict that timing and metacognitive performance should covary because both arise from the same accumulator dynamics. It equally challenges hierarchical accounts,[28,35] which predict that stronger timing encoding should produce better metacognitive judgment. The data were instead most consistent with a Higher-Order Representation account,[36,37] under which metacognitive evaluation requires the active re-representation of first-order timing states through mechanisms that are neurally distinct from those supporting primary task performance.

Taken together, the three studies build a consistent picture in which TEM is not merely a product of interval production signals, but rather an active, higher-order process operating on internally represented timing states. What is perhaps most theoretically significant is the convergence of evidence across levels of analysis: behavioral conservation across species established that the capacity exists independently of species-specific cognitive architecture; the motor demand dissociation established that it draws mostly on internal clock representations rather than peripheral feedback; and the neural decoding dissociation established that metacognitive evaluation recruits a representational organization that goes beyond the signals supporting timing performance, requiring coordinated multi-band integration rather than a direct readout of any single oscillatory substrate. Each study alone could be accommodated by multiple theoretical accounts. Together, they provide converging evidence against single-process and hierarchical accounts and are most consistent with a HOR framework, a position that itself aligns with broader proposals that conscious access to temporal information requires the conversion of automatically encoded neural dynamics into an abstract, evaluable form.[41] Within this broader landscape, TEM, as operationalized and investigated here, exemplifies temporal metacognition in its most directly measurable form: the capacity to read out and evaluate the reliability of one's own internal timing dynamics.

3.1) Limitations and Future Directions

Despite the consistent insights offered by the three studies, several limitations warrant consideration and point toward promising avenues for future research.

A first limitation concerns the cross-species comparison of motivational incentives. Human participants received symbolic monetary rewards, whereas rats received food pellets, incentive types that differ not only in modality but in the degree to which they engage immediate physiological versus abstract valuation processes. A skeptical reviewer might ask whether the cross-species strategy differences observed in Paper 1 reflect genuine cognitive differences in rule representation or simply differences in how urgently each species was motivated to perform accurately. While the theta alignment analyses partially address this by equating task-rule representation statistically, they cannot fully dissociate motivation from cognition. Future research could address this more directly by systematically varying reward magnitude and type within species, or by employing paradigms that more precisely equate motivational salience across species, for instance, through titration procedures that match subjective reward value.

A second limitation concerns the absence of neural measures in the cross-species and motor demand studies. Paper 2 established behaviorally that higher motor demands selectively impair metacognitive precision, supporting a clock-based account of TEM. However, the neural mechanism through which motor demand exerts this selective effect remains uncharacterized. One possibility consistent with the present findings is that increased motor variance degrades the signal-to-noise ratio of the internal clock representation at a neural level, reducing the quality of the metacognitive readout without affecting the central tendency of the interval estimate. An alternative possibility is that increased motor demands engage attentional resources that would otherwise support metacognitive evaluation, producing interference rather than signal degradation. These accounts generate distinct neural predictions that cannot be distinguished from behavioral data alone. The EEG study in Paper 3 was conducted under a single motor condition, leaving open whether the multi-band integration required for metacognitive inference is disrupted or reorganized when motor demands increase. Combining neural decoding with parametric motor demand manipulations in future work would allow a

direct test of these competing accounts and would strengthen the mechanistic bridge between the behavioral and neural findings of this thesis.

A third limitation concerns the EEG dataset size and its implications for the decoding analyses in Paper 3. After preprocessing, each participant contributed a relatively modest number of trials, which may be insufficient for Vision Transformer architectures that typically benefit from larger and more diverse training sets. This raises the question of whether the multi-band requirement observed for metacognitive inference reflects a genuine property of the underlying neural architecture or a sensitivity limitation of the decoding approach, that is, whether individual bands appear insufficient simply because there were too few trials to train reliable single-band classifiers. The fact that TP decoding succeeded reliably from individual bands within the same dataset provides partial reassurance that the multi-band requirement for MI is not merely a power artifact, since any global sensitivity limitation would have affected TP decoding equally. Nevertheless, replication with larger trial samples remains an important future step. Future studies may also benefit from combining deep learning decoding with hypothesis-driven oscillatory analyses to improve both reliability and interpretability of single-trial metacognitive decoding.

A fourth and broader direction concerns the informational architecture of multi-band integration itself. The present findings establish that no single frequency band was individually sufficient for metacognitive inference, whereas the simultaneous integration of θ, α, and β signals supported reliable decoding. Multi-band integration during ongoing motor action timing may reflect the simultaneous engagement of prediction, readout, and confidence estimation processes. Whether these functions are genuinely concurrent or reflect sequential processing within a single trial cannot be resolved from the present analyses. Future work systematically dissociating these band-specific contributions, for instance, through targeted perturbation or source-level analyses, will be essential for a complete mechanistic account of how the brain comes to know and evaluate the time it has made.

4) REFERENCES

1. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

2. Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

3. Paton, J. J., & Buonomano, D. V. (2018). The neural basis of timing: Distributed mechanisms for diverse functions. Neuron, 98(4), 687–705. https://doi.org/10.1016/j.neuron.2018.03.045

4. Foti, D., Kotov, R., Bromet, E., & Hajcak, G. (2012). Beyond the broken error-related negativity: functional and diagnostic correlates of error processing in psychosis. *Biological Psychiatry, 71*(10), 864–872. https://doi.org/10.1016/j.biopsych.2012.01.007

5. Endrass, T., Schuermann, B., Kaufmann, C., Spielberg, R., Kniesche, R., & Kathmann, N. (2010). Performance monitoring and error significance in patients with obsessive-compulsive disorder. *Biological Psychology, 84*(2), 257–263. https://doi.org/10.1016/j.biopsycho.2010.02.002

6. Buonomano, D. V., & Maass, W. (2009). State-dependent computations: Spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience*, *10*(2), 113–125. https://doi.org/10.1038/nrn2558

7. Gibbon, J. (1977). Scalar expectancy theory and Weber's law in animal timing. *Psychological Review, 84*(3), 279–325. https://doi.org/10.1037/0033-295X.84.3.279

8. Simen, P., Balci, F., DeSouza, L., Cohen, J. D., & Holmes, P. (2011). Interval timing by long-range temporal integration. *Frontiers in Integrative Neuroscience, 5*, 28. https://doi.org/10.3389/fnint.2011.00028

9. Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026. https://doi.org/10.1038/nn.2590

10. Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*(11), 556–564. https://doi.org/10.1016/j.tics.2013.09.009

11. Akdoğan, B., & Balcı, F. (2017). Are you early or late?: Temporal error monitoring. *Journal of Experimental Psychology: General, 146*(3), 347–361. https://doi.org/10.1037/xge0000265

12. Kononowicz, T. W., & van Wassenhove, V. (2019). Evaluation of self-generated behavior: Untangling metacognitive readout and error detection. *Journal of Cognitive Neuroscience, 31*(11), 1641–1657. https://doi.org/10.1162/jocn_a_01442

13. Öztel, T., & Balcı, F. (2023). Humans can monitor trial-based but not global timing errors: Evidence for relative judgements in temporal error monitoring. *Quarterly Journal of Experimental Psychology, 76*(9), 2155–2163. https://doi.org/10.1177/17470218221145314

14. Öztel, T., & Balcı, F. (2024). Metric error monitoring as a component of metacognitive processing. *The European Journal of Neuroscience, 59*(5), 807–821. https://doi.org/10.1111/ejn.16182

15. Kononowicz, T. W., van Wassenhove, V., & Doyère, V. (2022). Rodents monitor their error in self-generated duration on a single-trial basis. *Proceedings of the National Academy of Sciences, 119*(9), e2108850119. https://doi.org/10.1073/pnas.2108850119

16. Balcı, F., Freestone, D., & Gallistel, C. R. (2009). Risk assessment in man and mouse. *Proceedings of the National Academy of Sciences of the United States of America, 106*(7), 2459–2463. https://doi.org/10.1073/pnas.0812709106

17. Balci, F., Freestone, D., Simen, P., DeSouza, L., Cohen, J. D., & Holmes, P. (2011). Optimal temporal risk assessment. *Frontiers in Integrative Neuroscience, 5*, 56. https://doi.org/10.3389/fnint.2011.00056

18. Foote, A. L., & Crystal, J. D. (2007). Metacognition in the rat. *Current Biology, 17*(6), 551–555. https://doi.org/10.1016/j.cub.2007.01.061

19. Jozefowiez, J., Staddon, J. E. R., & Cerutti, D. T. (2009). Metacognition in animals: How do we know that they know? *Comparative Cognition & Behavior Reviews, 4*, 29–39. https://doi.org/10.3819/ccbr.2009.40003

20. Bilgin, S. N., Doyère, V., & Kononowicz, T. W. (2026). Direct comparison of temporal error monitoring in humans and rats. *iScience*, 29(1), 115531. https://doi.org/10.1016/j.isci.2026.115531

21. Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics, 14*(1), 5–12. https://doi.org/10.3758/BF03198607

22. Bilgin, S. N., & Kononowicz, T. W. (2025). Temporal error monitoring: Monitoring of internal clock or just motor noise? Consciousness and Cognition, 130, 103849. https://doi.org/10.1016/j.concog.2025.103849

23. Kononowicz, T. W., & van Rijn, H. (2015). Single-trial beta oscillations index time estimation. *Neuropsychologia, 75*, 381–389. https://doi.org/10.1016/j.neuropsychologia.2015.06.014

24. Kulashekhar, S., Pekkola, J., Palva, J. M., & Palva, S. (2016). The role of cortical beta oscillations in time estimation. *Human Brain Mapping, 37*(9), 3262–3281. https://doi.org/10.1002/hbm.23239

25. Engel, A. K., & Fries, P. (2010). Beta-band oscillations—signalling the status quo? Current Opinion in Neurobiology, 20(2), 156–165. https://doi.org/10.1016/j.conb.2010.02.015

26. Spitzer, B., & Haegens, S. (2017). Beyond the status quo: A role for beta oscillations in endogenous content (re)activation. eNeuro, 4(4), ENEURO.0170-17.2017. https://doi.org/10.1523/ENEURO.0170-17.2017

27. Tan, H., Wade, C., & Brown, P. (2016). Post-movement beta activity in sensorimotor cortex indexes confidence in the estimations from internal models. Journal of Neuroscience, 36(5), 1516–1528. https://doi.org/10.1523/JNEUROSCI.3204-15.2016

28. Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B*, 367(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

29. Kononowicz, T. W., Roger, C., & van Wassenhove, V. (2019). Temporal metacognition as the decoding of self-generated brain dynamics. *Cerebral Cortex, 29*(10), 4366–4380. https://doi.org/10.1093/cercor/bhy318

30. Barne, L. C., Claessens, P. M. E., Reyes, M. B., Caetano, M. S., & Cravo, A. M. (2017). Low-frequency cortical oscillations are modulated by temporal prediction and temporal error coding. NeuroImage, 146, 40–46. https://doi.org/10.1016/j.neuroimage.2016.11.028

31. Balcı, F., & Simen, P. (2016). A decision model of timing. Current Opinion in Behavioral Sciences, 8, 94–101. https://doi.org/10.1016/j.cobeha.2016.02.002

32. Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. Wiley.

33. Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 443. https://doi.org/10.3389/fnhum.2014.00443

34. Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

35. Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*(1), 91–114. https://doi.org/10.1037/rev0000045

36. Rosenthal, D. M. (1986). Two concepts of consciousness. Philosophical Studies, 49(3), 329–359. https://doi.org/10.1007/BF00355521

37. Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. Trends in Cognitive Sciences, 23(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

38. Bilgin, S. N., Giomo, D., Mustafali, U., & Kononowicz, T. W. (2026). *The brain as actor and evaluator: Distinct neural codes for timing and self-evaluation of timing errors revealed by transformer-based decoding* [Preprint]. bioRxiv. https://doi.org/10.64898/2026.03.24.714009

39. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition

at scale. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.2010.11929

40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

41. van Wassenhove, V. (2009). Minding time in an amodal representational space. *Philosophical Transactions of the Royal Society B, 364*(1525), 1815–1830. https://doi.org/10.1098/rstb.2009.0023

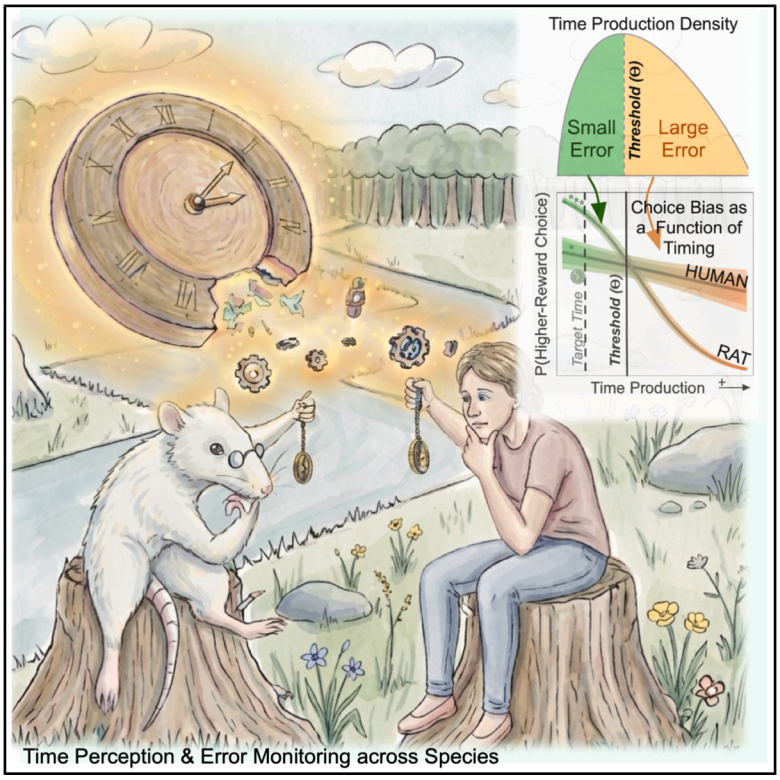
5.1) Paper 1 : *Direct Comparison of Temporal Error Monitoring in Humans and Rats*

iScience

Article

Direct comparison of temporal error monitoring in humans and rats

Graphical abstract



Highlights

- Rats and humans can report their own temporal errors in a time production task
- Rats show higher temporal error monitoring accuracy than humans
- Proximity to the decision threshold improves error-reporting accuracy
- Response times reveal confidence differences between species

Authors

Sena N. Bilgin, Valérie Doyère, Tadeusz W. Kononowicz

Correspondence

sbilgin@sd.psych.pan.pl (S.N.B.), tadeusz.kononowicz@cnrs.fr (T.W.K.)

In brief

Neuroscience; Techniques in neuroscience; Psychology

Bilgin et al., 2026, iScience *29*, 115531
May 15, 2026 © 2026 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2026.115531

CellPress

iScience

CellPress
OPEN ACCESS

Article

Direct comparison of temporal error monitoring in humans and rats

Sena N. Bilgin,[1,3,*] Valérie Doyère,[2] and Tadeusz W. Kononowicz[1,2,*]
[1]Institute of Psychology, Polish Academy of Sciences, ul. Jaracza 1, 00-378 Warszawa, Poland
[2]Paris-Saclay University, CNRS, Paris-Saclay Institute of Neuroscience (NeuroPSI), 91400 Saclay, France
[3]Lead contact
*Correspondence: sbilgin@sd.psych.pan.pl (S.N.B.), tadeusz.kononowicz@cnrs.fr (T.W.K.)
https://doi.org/10.1016/j.isci.2026.115531

SUMMARY

Temporal error monitoring sits at the intersection of timing, metacognition, and error monitoring. Through a comparative paradigm, we explored distinctions and parallels between rats and humans. Both species were trained to produce a target interval (2s) by pressing a key and to associate reward size based on error magnitude with the location of one of two cards on a screen (humans) or with reward ports (rats). Results showed that both species could report their temporal errors, with rats displaying higher accuracy. Better representation of rule-based variables increased choice accuracy in both species. When performances were equated for specific task variables, the accuracy difference diminished but remained. Response times, used as a proxy for confidence, reflected confidence more accurately in rats, likely resulting from different exploratory and decisional processes in humans. These findings identify key task variables that support temporal error monitoring, with broader implications for comparative metacognition research.

INTRODUCTION

The comparative and translational examination of cognitive processes, including metacognition and performance monitoring, can provide insights into the extent to which these processes are similar across species. This type of study involves designing experiments, training humans and rats to perform specific tasks, and observing their behavior to understand their cognitive abilities. However, it is crucial in designing experiments to consider both construct and computational validity.[1] Construct validity seeks to answer whether we are measuring the same theoretical construct across different species. For example, if experimenters measure reaction time, it should be defined in a similar way in the considered species. However, when training rodents, can we be confident that they are performing tasks in a manner comparable to humans? Computational/algorithmic validity concerns the representation and internalization of task-relevant rules, as well as computations performed on those representations. It acknowledges that species-specific or shared (similar) behaviors may emerge from different underlying algorithmic processes. Needless to say, the discovery of the same or shared algorithmic processes often encounters obstacles such as variations in design details and differences in the understanding of task rules between human and non-human species, which often lead to controversial results. In particular, the interpretation of findings from studies on metacognition and performance monitoring across species remains a matter of ongoing debate. The key challenge is the scarcity of studies directly comparing humans and rats under the same experimental conditions. To address this, the present study adapted a paradigm originally developed in rats[2] for humans to enable a systematic cross-species comparison of putative underlying metacognitive processes as expressed through behavioral performance at the single-trial level.

Previous studies highlight the debate on the extent of these cognitive abilities within the species under investigation.[3–6] For example, in the temporal discrimination task conducted by Foote and Crystal,[7] rats displayed the ability to assess and indicate their uncertainty in duration discrimination. In trials close to the discrimination threshold, rats expressed their uncertainty by favoring safer bets instead of riskier ones with a long time-out punishment. This insight into rats' metacognitive processes suggests their capacity to introspectively evaluate the reliability of their temporal judgments. Although similar results have not been questioned in humans,[8,9] low-level interpretations, using various associative models, have been proposed to account for the uncertainty-monitoring data produced by rodents. To fit the uncertainty-monitoring data produced by rats in the temporal discrimination task used in Foote and Crystal,[7] Jozefowiez[10] employed a signal-detection model, suggesting that animals may have used different task variables to produce the pattern of uncertain opt-out responses.

The study by Foote and Crystal[7] serves as an incisive task to study timing and uncertainty together. Recent studies pushed the envelope further, discovering that humans can keep track of errors and uncertainty in their own timing behavior.[11,12] In these studies, combining timing and performance monitoring, it has been observed that individuals, when instructed to produce or reproduce a specific time interval, consistently report the extent of their timing errors, as well as the direction of these errors.

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Going beyond humans, there has been a growing interest in questioning whether rodents share the ability to monitor timing errors relative to target durations on a trial-by-trial basis. Kononowicz et al.[2] demonstrated that rats can accurately report the magnitude of their just-produced timing errors by indicating how far off they were from the target duration (also see Oztel and Balcı[13]). A key aim of the present study was to directly compare performance and task rule utilization in error monitoring across rats and humans by applying a closely matched experimental design. To compare performance and task rule utilization in error monitoring in rats and humans, we used the paradigm we had developed in rats[2] and adapted it for human testing. Both species were trained to produce a target time interval (2s) and then report the magnitude of their error immediately after each time production (TP). The task followed three phases: duration production training, temporal error monitoring training, and temporal error monitoring testing, which were adapted to humans but closely paralleled the procedure used in rats. Across these phases, humans and rats learned to reproduce the target interval under probabilistic reward contingencies and later to make choices that reflected their perceived temporal errors. This structure provided a unified behavioral framework for examining how both species use task rules to guide error reporting. By aligning the methodological procedures across species, the study aimed to enable a robust comparison of behavioral strategies in temporal error monitoring. Although we do not claim to identify shared neural substrates or specific computational or metacognitive mechanisms, the cross-species behavioral profiling and extraction of rule-based variables could reveal converging patterns in how the task is approached. Establishing such a comparable behavioral framework provides a necessary foundation for interpreting future research integrating neural recordings or computational modeling to investigate potential mechanistic commonalities across species further.

RESULTS

Rats and humans produce time intervals optimally

Humans and rats were trained to produce a minimum target duration (T = 2 s) by initiating an interval with a brief motor action (lever or key press) and ending it with a second action once they judged that the target time had elapsed (Figure 1A). Two experimental factors were manipulated: the motor action used to produce the interval and the explicitness of task instructions. This yielded PRESS and HOLD groups in both humans and rats, in which the interval was indicated by a second press or terminated by releasing the initiating press, respectively, allowing us to test whether different motor routines influenced temporal error monitoring. Instruction explicitness was varied in humans to better match the rat task experience, resulting in instruction (inst) and reduced-instruction (no-inst) groups. Together, this design produced four groups in humans: PRESS/inst, PRESS/no-inst, HOLD/inst, and HOLD/no-inst. Data were pooled across conditions within each species for all subsequent analyses, as no condition differences were observed in key performance metrics (see Figures S4 and S5).

Producing at least the target time (T) gave humans and rats access to a reward. All groups produced the required time interval T (Figure 1B), with distributions of time productions (TPs) peaking after T. Rats demonstrated higher accuracy (Figure 1B, left panel; $\beta = -0.10$, 95% CI [−0.15, −0.05], t (442) = −4.04, $^{***}p < 0.001$) and precision (Figure 1B, right panel; $\beta = -0.05$, 95% CI [−0.08, −0.02], t (442) = −3.51, $^{***}p < 0.001$) in TPs compared to humans. Rats' mean TPs (μ[TPs]) across sessions were closer to the target duration than those of humans. Despite the difference in precision and accuracy between species, the standard deviation of TPs (σ[TPs]) increased with the mean TPs (μ[TPs]) similarly in both species (Figure 1B, right panel inset plot; $\beta = 0.23$, 95% CI [−0.08, 0.54], t (442) = 1.47, $p = 0.143$), following Weber's law in duration reproduction.[14–16]

Overall, these findings indicate that while there are species-specific differences in the accuracy and precision of time production, both humans and rats exhibit similar lawful variability in their time production, following Weber's law.

Rats and humans express error monitoring

A previous study found that rats were able to track their ongoing TPs, influencing their choices.[2] Therefore, we assessed whether the current-trial TP is a factor in decision-making in both species. We first fitted a GLMM to predict the probability of choosing a small error (SE) port/card based on the current trial TP across humans/rats and sessions (Figure 1D). In a global analysis, we confirmed that the probability of choosing an SE port/card decreased as a function of the current-trial TP (Figure S1A, μTPn; $\beta = -0.79$, 95% CI [−0.82, −0.75], $^{***}p < 0.001$), and as a function of previous trials' TP errors (Figure S1B, μTP(10); $\beta = -1.00$, 95% CI [−1.07, −0.93], $^{***}p < 0.001$). The relationship between current-trial performance and the probability of choosing the SE port/card was found to be stronger in rats (Figure 1D, left panel, rats (blue line); $\beta = -0.99$, 95% CI [−1.05, −0.92], $^{***}p < 0.001$) than in humans (Figure 1D, left panel, humans (red line); $\beta = -0.13$, 95% CI [−0.24, −0.02], $^{*}p = 0.017$). Adding a "species" factor to the baseline model was justified (ΔAIC = 182, $^{***}p < 0.001$). This suggests that rats relied more on current-trial TPs than humans when making their choices during test trials ($\beta = -0.86$, 95% CI [−0.98, −0.73], $^{***}p < 0.001$). In other words, while both species utilized current-trial TPs in decision-making, rats showed a stronger reliance on these TPs, whereas humans relied more on previous trials compared to rats (Figure S1C, $\beta = 0.25$, 95% CI [0.02, 0.48], $^{*}p = 0.030$).

Since the reward rule was dependent on the size of the TP error, test trial accuracy (the proportion of correct choices) is a direct measure of temporal error monitoring. From the right panel of Figure 1D, it is evident that accuracy on test trials was above the chance level for both humans (red, 95% CI [0.517, 0.574], $^{**}p = 0.002$, two-sided Wilcoxon test) and rats (blue, 95% CI [0.645, 0.668], $^{***}p < 0.001$, two-sided Wilcoxon test). However, rats were more accurate in their choices than humans (Figure 1D, right panel; $\beta = 0.09$, 95% CI [0.04, 0.14], t (441) = 3.40, $^{***}p < 0.001$), indicating significant differences between species in terms of choice accuracy during test trials.

Overall, when examining temporal error monitoring, interesting dynamics were observed in both species. Rats exhibited a stronger influence of TP errors on their choices, indicating a more robust

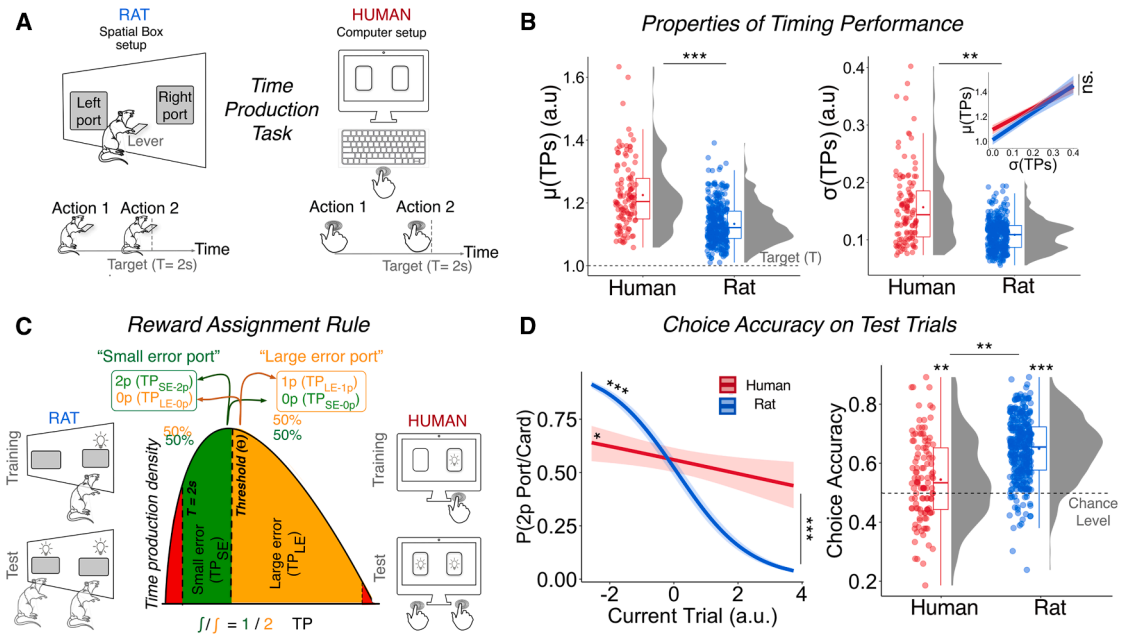


Figure 1. Rats and humans produce time intervals optimally and report temporal errors by choosing accurately on test trials

(A) Humans and rats were trained on the time production task. Humans and rats were trained on a time production task. Depending on group assignment, they either held down a lever/key for a minimum target duration and released it (HOLD group) or pressed the lever twice with the required delay (PRESS group). For rats, it involved reward ports on either side of a box; for humans, the task involved choosing between reward cards on the left and right sides of a screen. The location of a reward accessibility was indicated by illuminating the corresponding card or port if the participant's and rats' time production exceeded the target. Humans received rewards by selecting the illuminated card, while rats did so by nose-poking the illuminated reward port.

(B) Time production (TP) performance in error monitoring test sessions. Boxplots with density curves illustrate accuracy (mean, μ) and precision (standard deviation, σ) of TP for rats (blue) and humans (red) at the session level. Data are presented in arbitrary units (a.u.) to enable comparison across conditions. Rats showed higher accuracy, with μ(TPs) closer to the target duration, and higher precision, reflected in smaller σ(TPs). However, σ(TPs) increased with μ(TPs) similarly in both species, indicating conserved timing response properties. Statistical significance is indicated by ***, **, *, ns for $p < 0.001$, 0.01, and 0.05, and $p > 0.05$, respectively. Confidence intervals, shown as shaded areas, were plotted for each condition at population level. The width of these shaded regions reflects population level estimates, taking into account uncertainty in species parameter.

(C) Diagram illustrating the reward assignment rule during error monitoring training and test trials. The arrows indicate the probabilistic assignment of TP type (small error (SE), colored green or large error (LE), colored orange) to left and right cards/ports on training trials. During the training trials, all humans and rats were trained to associate a reward size and the location of one of two presented cards/ports with the error magnitude in their time production. The assignment ensured a 50% reward probability upon selecting the illuminated option. In test trials, reward-location associations remained, but both options were available for choice.

(D) Choice accuracy on test trials. GLMMs predicted the probability of choosing the 2p (SE) port/card from the current-trial TP. Shorter TPs increased the likelihood of 2p choice, indicating single-trial temporal error monitoring. Rats showed steeper slopes than humans, suggesting stronger reliance on current TP. Confidence intervals reflect group-level variance. Accuracy (right panel) was significantly above chance (0.5) for both species, with rats showing higher overall accuracy.

temporal error monitoring. This difference might contribute to the striking contrast in choice accuracy between the two species.

Comprehension of task variables improves temporal error monitoring across species

The ability to perform well in temporal error monitoring was hypothesized to increase with a precise representation of task variables. Humans and rats ought to keep track of the value assigned to them based on their distribution of TPs in prior sessions, referred to as "assigned" theta, which governs the reward assignment rule. The second variable, "observed" theta, is the value computed based on the reward assignment rule considering the TP distribution in the current session. Lastly, the "choice" theta is derived by examining humans' and rats' decisions regarding SE and large error (LE) choices within the context of the TP distribution in the current session. Essentially, we calculated the point of subjective equivalence between (SE) and (LE) choices.

At the session level, calculations were made to assess the relationship between the three theta values by taking into account the distance between them. We hypothesized that smaller difference between theta values would indicate better theta alignment, leading to a more precise context representation and thus improved accuracy on test trials. To test this hypothesis, session-based choice accuracy values were correlated with three indices of distances between theta distributions. GLMMs, with each index as a predictor of choice accuracy values on a per-session basis across humans and rats, were fitted. As the humans and rats aligned the task variables more effectively (decreased theta alignment), session-level selection accuracy values were found to be higher ($\beta = -0.06$, 95% CI [−0.07, −0.04], $t(1250) = -9.23$, $^{***}p < 0.001$).

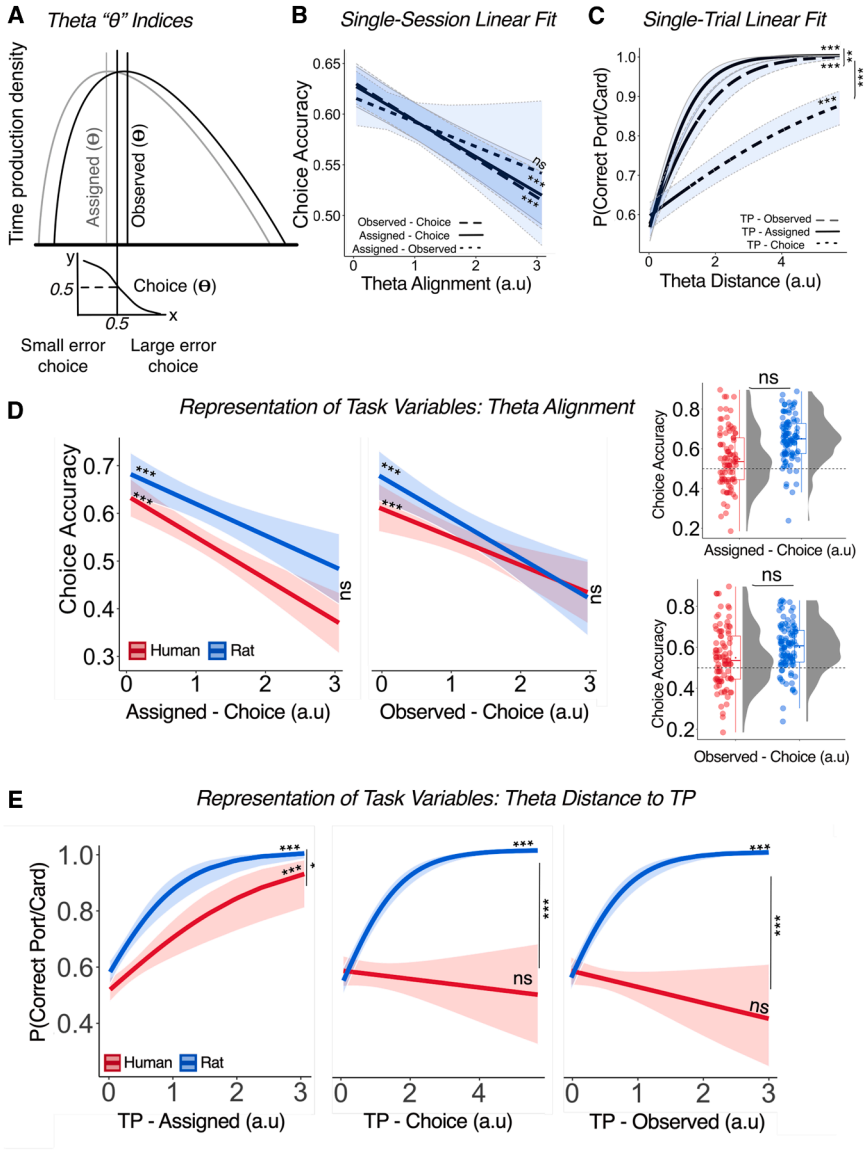


Figure 2. Similarities and differences in comprehension of task variables across species

(A) Theta "θ" indices as a representation of task variables. Three theta values captured different task aspects: "assigned" theta: the experimenter-assigned value based on prior TP distributions for reward rules; "observed" theta: calculated via the reward rule using current session TP data; and "choice" theta: the decision boundary where humans and rats cannot distinguish small vs. large errors.

(B) Session-based task variables. At the session level, the relationship between the three different theta values was calculated to understand how well humans and rats aligned the task variables to create the representation of task contexts. GLMMs modeled each theta alignment index as a predictor of choice accuracy per session. The "observed—choice" and "assigned—choice" indices strongly correlated with accuracy, whereas "assigned—observed" showed marginal correlation. Values were in arbitrary units to enable cross-condition comparison. Statistical significance is indicated by ***, **, *, ns for $p < 0.001$, 0.01, and 0.05, and $p > 0.05$, respectively.

(C) Single-trial-based task variables. At the trial level, task variables were distances between the current trial TP and each theta value. GLMMs predicted the probability of correct choices based on these distances. Larger distances corresponded to higher probabilities of correct choices, indicating humans and rats monitored temporal errors on a single-trial basis and chose accordingly when TP was far from the relevant theta distribution.

(D) Choice accuracy as a function of Theta "θ" alignments across species. To ensure valid comparisons between species, each θ value session of the human group was matched to the closest rat data points. We then assessed whether the choice accuracy of the different species, sampled according to this index, would differ. GLMMs showed choice accuracy significantly correlated with theta alignment indices in both species, with no significant differences. This indicates similar performance when session-level task representation was matched across species.

(E) Probability of choosing the correct port as a function of TP distance to Theta indices across different species. GLMMs examined the correct choice probability as a function of TP distance to theta indices per species. In rats, all distances positively predicted correct choices. In humans, only the distance to "assigned" theta correlated positively. Species interactions suggest rats more effectively used trial-by-trial TP positioning relative to all theta values to guide choices than humans.

To explore the relationship between each theta index and session-based accuracy, we predicted accuracy using theta alignment indices on a per-session basis across humans and rats. Session-based choice accuracy values were correlated with all three theta alignment indices separately, and the interaction between these indices was tested. Notably, the "observed—choice" index (Figure 2B, dashed line; $\beta = -0.06$, 95% CI [−0.08, −0.04], t (1246) = −7.29, ***$p < 0.001$) and the "assigned—choice" index (Figure 2B, solid line; $\beta = -0.06$, 95% CI [−0.07, −0.04], t (1246) = −6.61, ***$p < 0.001$) exhibited a robust correlation with humans' and rats' choice accuracy at the session level. However, a comparable relationship was not observed between the "assigned—observed" index and choice accuracy values (Figure 2B, dotted line; $\beta = -0.04$, 95% CI [−0.08, 7.91e−03], t (1246) = −1.62, $p = 0.105$). These results led to the exclusion of this latter index for the subsequent analyses.

Discriminability of temporal error depends on distance from threshold

We hypothesized that both species should keep track of the distance between the three distinct theta values and the current trial

TP. Larger distances between TPs and theta values were hypothesized to enhance discriminability on a trial-by-trial basis. To test this hypothesis, single-trial accuracy values were correlated with each distance index. A GLMM with each distance index as a predictor of the probability of selecting the correct port/card on a trial basis across humans and rats and sessions was fitted. As the distance of theta to the current TP increased, the probability of selecting the correct port/card was found to be higher ($\beta = 0.37$, 95% CI [0.31, 0.43], $^{***}p < 0.001$). Moreover, significant differences between the slopes of TP-theta distances were observed ($\Delta AIC = 108$, $^{***}p < 0.001$). This suggests their ability to monitor temporal errors on a single-trial basis and make accurate choices when the current TP diverges significantly from the relevant theta. A steeper slope of the logistic function indicated an elevated probability of accurately choosing the correct port, corresponding to a greater distance from the associated theta value, for both "assigned" (Figure 2C, solid line; $\beta = 1.40$, 95% CI [1.13, 1.67], $^{***}p < 0.001$) and "observed" (Figure 2C, dashed line; $\beta = 0.97$, 95% CI [0.78, 1.16], $^{***}p < 0.001$). The "choice" (Figure 2C, dotted line; $\beta = 0.28$, 95% CI [0.21, 0.35], $^{***}p < 0.001$) had a less pronounced influence compared to other distance indices, with the steepest slope belonging to the "assigned" index.

Equating for task variables reveals equivalent accuracy between species

As we established that the representation of task variables improves choice accuracy, we then performed cross-species comparisons while accounting for potential differences in task representation. To that end, the data point of each theta index (comprising session values) from the human group was matched with the closest corresponding rat data point (see Figure S2). This procedure allowed for an assessment of whether the choice accuracies from each species aligned according to these indices were consistent. We conducted two types of analyses: one using choice accuracy and the other relying on estimates from the regression model, which includes the effects of previous trials.

Next, we assessed the correlation of indices of representation of task variables with choice accuracy. For both species, we found significant associations for "observed−choice" (Figure 2D, right panel, rats (blue line); $\beta = -0.10$, 95% CI [−0.14, −0.05], $t(176) = -4.47$, $^{***}p < 0.001$; humans (red line); $\beta = -0.07$, 95% CI [−0.11, −0.03], $t(176) = -3.52$, $^{***}p < 0.001$) and "assigned−choice" indices (Figure 2D, left panel, rats (blue line); $\beta = -0.08$, 95% CI [−0.11, −0.04], $t(176) = -3.97$, $^{***}p < 0.001$; humans (red line); $\beta = -0.10$, 95% CI [−0.13, −0.07], $t(176) = -5.88$, $^{***}p < 0.001$). Notably, no significant slope variations for both the "observed−choice" (Figure 2D, right panel, $\beta = -0.03$, 95% CI [−0.09, 0.03], $t(176) = -1.02$, $p = 0.295$) and the "assigned−choice" indices (Figure 2D, left panel, $\beta = 0.02$, 95% CI [−0.03, 0.07], $t(176) = 0.93$, $p = 0.351$) were detected, suggesting comparable ways of task variable representation. Using the estimates of Current TP instead of Choice Accuracy confirmed qualitatively the same results, such that current TP estimates correlated with task variable representation with the same strength (see Figure S3). The lack of interaction effect suggests that humans and rats represented the task structure similarly. Notably, in this portion of the data aligned by task representation, the accuracy gap between species disappeared. We did not find any significant effect of species for "observed−choice" (Figure 2D, bottom inset panel; $\beta = 0.08$, 95% CI [−5.75e−03, 0.16], $t(176) = 1.83$, $p = 0.068$) and "assigned−choice" indices (Figure 2D, upper inset panel; $\beta = 0.05$, 95% CI [−0.01, 0.12], $t(176) = 1.56$, $p = 0.121$). In sum, after equating the species in terms of task variables, choice accuracy was similar in both species, suggesting that they utilize task variables similarly.

Rats and humans use different task variables on single trials

In addition to the representation and alignment of different thresholds, the species could differ concerning which properties of temporal distributions they base their choices on. As we already verified that the discriminability of temporal error depends on the distance from the threshold, we set out to test how it differs across species. We predicted the likelihood that the species would differ in the task parameter they base their choice on at the level of individual trials, assessing the TP distance to "assigned," "observed," and "choice" parameters. Adding the "species" factor demonstrated significantly greater explanatory power for the variability in the data compared to the baseline models of each theta, as evidenced by the ΔAIC values ("choice": $\Delta AIC = 48$, $^{***}p < 0.001$; "observed": $\Delta AIC = 79$, $^{***}p < 0.001$; "assigned": $\Delta AIC = 13$, $^{***}p < 0.001$). Our global analysis revealed a positive slope in the logistic function, indicating that, as the distance of TPs to the relevant theta value increased, so did the probability of selecting the correct port/card on a trial-by-trial basis.

In rats (Figure 2E, blue line), all distance indices showed a positive correlation with the probability of making the correct choice (Figure 2E, middle panel, ("choice"): $\beta = 1.11$, 95% CI [0.49, 1.73], $^{***}p < 0.001$; rightmost panel, ("observed"): $\beta = 2.03$, 95% CI [1.59, 2.48], $^{***}p < 0.001$; leftmost panel, ("assigned"): $\beta = 1.58$, 95% CI [1.07, 2.10], $^{***}p < 0.001$, respectively). Conversely, in humans (Figure 2E, red line), this correlation was observed solely with the TP distance to the "assigned" parameter, where a significant positive association was found (Figure 2E, leftmost panel, ("assigned"): $\beta = 0.79$, 95% CI [0.41, 1.17], $^{***}p < 0.001$). No significant associations were found between the TP distance to "choice" or "observed" and the probability of making the correct choice (Figure 2E, middle panel, ("choice"): $\beta = -0.06$, 95% CI [−0.22, 0.10], $p = 0.473$; rightmost panel, ("observed"): $\beta = -0.22$, 95% CI [−0.52, 0.07], $p = 0.143$). Furthermore, the interactions between species were evident across all models. Notably, compared to humans, the rats displayed the most adept monitoring of their current TP relative to all theta distributions, adjusting their choices accordingly (Figure 2E, middle panel ("choice"): $\beta = 1.17$, 95% CI [0.59, 1.75], $^{***}p < 0.001$; rightmost panel ("observed"): $\beta = -2.26$, 95% CI [−2.79, −1.72], $^{***}p < 0.001$; leftmost panel ("assigned"): $\beta = 0.79$, 95% CI [0.16, 1.43], $^{*}p = 0.014$).

Overall, while rats and humans learned the same task structure, it appears that rats were more precise at representing the distribution of TPs in the current session and their associated errors on single trials. In contrast, humans seemed to rely more on the memory of previous performance and exhibited lower choice

accuracy. Therefore, the two species learned the same task structure but diverged in their utilization of task variables.

Response times on test trials provided insight into choice behavior

To examine whether humans and rats exhibited choice behavior, response times (RTs) on test trials per human and rat were compared across species and contrasted with training trials (the test phase encompasses both training and test trials; see the STAR Methods section for details). Significant RT differences between trial types were observed in both groups: for humans (Figure 3A, (red); $\beta = 0.78$, 95% CI [0.67, 0.89], t (96) = 13.91, $^{***}p < 0.001$) and for rats (Figure 3A, (blue); $\beta = 0.39$, 95% CI [0.23, 0.56], t (96) = 4.84, $^{***}p < 0.001$). Interestingly, humans exhibited a more pronounced RT difference compared to rats (Figure 3A; $\beta = -0.38$, 95% CI [−0.58, −0.19], t (96) = −3.88, $^{***}p < 0.001$). With RTs' difference being larger for humans, this suggests a different way of responding on test trials compared to rats. Several factors could play a role, ranging from different patterns of explorative behavior between species to differences in decision-making processes.

Both rats and humans exhibited faster responses when anticipating larger subjective internally based reward prospects

During the training trials of the test phase (excluding test trials), where only one port was illuminated, both rats (Figure 3B, (blue); $\beta = 0.18$, 95% CI [0.14, 0.22], t (96) = 8.72, $^{***}p < 0.001$) and humans (Figure 3B, (red); $\beta = 0.05$, 95% CI [0.02, 0.08], t (96) = 3.72, $^{***}p < 0.001$) reached or chose the port/card faster when anticipating reward delivery (R) compared to non-rewarded trials (NR), based on their TP error. Notably, the magnitude of this difference was more pronounced in rats ($\beta = 0.12$, 95% CI [0.08, 0.17], t (96) = 5.09, $p < 0.001$), highlighting the influence of reward-driven processes on RTs.

RTs as an indicator of confidence

RT (RT) can provide valuable insights into decision-making processes and the associated confidence levels with a task.[17] Here, we utilize RTs to characterize the processes guiding decisions and associated confidence in the context of temporal error monitoring. Such an approach has been proven valid in the context of temporal error monitoring across species.[2,18] If decisions in temporal error monitoring can be described in terms of the accumulation of evidence, one would predict that: (1) accurate trials should be associated with faster RTs when the expectation of momentary evidence is strong, and (2) for trials where the rate of evidence is high, RTs should be shorter.

At the trial level, we explored the relationship between RTs and the probabilities of making a correct choice across species. The inclusion of the "species" factor in the model was justified (ΔAIC = 12, $^{***}p < 0.001$) compared to the base model, which solely associates RTs with the probability of choosing the correct port. The interaction between species and RTs was significant (Figure 3C; $\beta = -0.18$, 95% CI [−0.30, −0.06], $^{**}p = 0.003$). The effect of RTs was statistically significant only in rats, indicating that their confidence indices were still discernible on single trials (Figure 3C, rats (blue line); $\beta = -0.23$, 95% CI [−0.33,

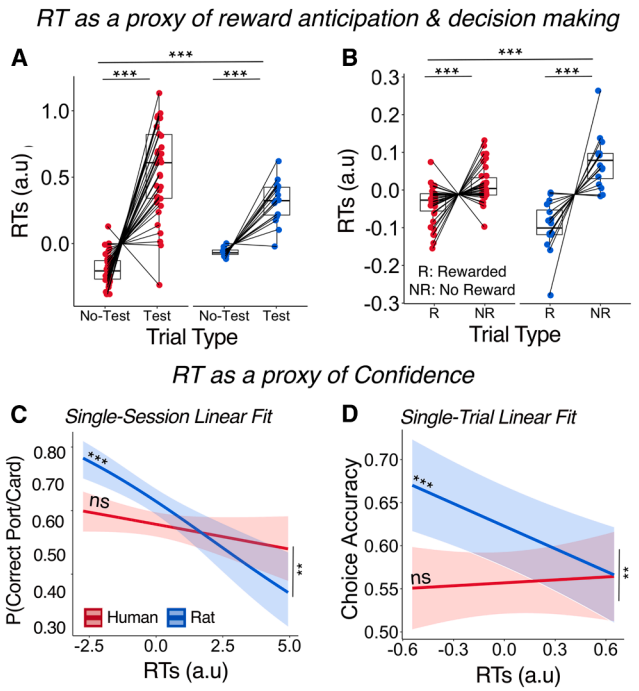


Figure 3. Both rats and humans exhibited faster choice responses when anticipating larger subjective internally based reward prospects

(A) Response times (RTs) as an indicator of choice behavior. During training, humans and rats were trained to associate their time production (TP) error with either a small error (SE) or large error (LE) port/card. In test trials, both options were illuminated, requiring humans and rats to choose. Both species received a reward if they selected the port/card corresponding to the magnitude of the error in their TP at that trial. Response time (RT), the time from illumination to nose poking (rats) or card selection (humans), was analyzed as a measure of decision-making. RTs were z-scored to allow cross-species comparison (see data processing). RTs were longer in test (choice) trials than training (no-choice) trials across both species, with humans (red) showing a greater increase than rats (blue), suggesting additional processing during test trials. Significance levels are marked as ***, **, *, ns for $p < 0.001$, 0.01, and 0.05, and $p > 0.05$, respectively.

(B) Response times (RTs) as an indicator of reward anticipation. Consistent with the previous study,[2] during the training trials of the test phase, both species reached/chose the port/card faster when anticipating reward delivery based on their TP error. This effect was more pronounced in rats (blue) compared to humans (red).

(C and D) Response times (RTs) as an indicator of confidence. To explore the significance of response times (RTs) as markers of confidence across various scenarios, we investigated RTs (normalized by the z-scoring method; see data processing for details) both within single trials and sessions. In (C), RTs (x axis) were subjected to analysis via a GLMM, serving as a predictor of the probabilities of selecting the correct port/card on a trial-by-trial basis. Rats are depicted in blue, while humans are represented in red. RTs significantly correlated with the probability of choosing correctly solely in rats. The observed interaction between species and RTs indicated the presence of trial-level confidence exclusively in rats. In (D), utilizing a GLMM, normalized RTs were examined as predictors of choice accuracy values at the single-session level. Notably, while RTs significantly correlated with choice accuracy in rats, this association was absent in humans. This disparity emphasizes distinct session-level confidence indicators between species.

−0.14], ***$p < 0.001$; humans (red line); $\beta = -0.05$, 95% CI [-0.12, 0.03], $p = 0.202$).

To assess whether RT serves as a proxy for confidence in a straightforward manner, we evaluated the relationship between RTs and choice accuracy at the single-session level. The inclusion of the "species" factor in the model was justified (ΔAIC = 12.9, ***$p < 0.001$) compared to the base model, which only associates RTs with single-session choice accuracy values. The interaction effect between species and RTs was significant (Figure 3D; $\beta = -0.12$, 95% CI [−0.19, −0.05], t(903) = −3.29, **$p = 0.001$). An RT-based signature of confidence was found only in rats (Figure 3D, blue line; $\beta = -0.10$, 95% CI [-0.15, −0.06], t(903) = −4.57, ***$p < 0.001$), while no RT effects of confidence were found in humans (Figure 3D, red line; $\beta = 0.02$, 95% CI [-0.04, 0.08], t(903) = 0.64, $p = 0.523$). Overall, in sessions with higher choice accuracy and trials where the probability of selecting the correct port was greater, rats exhibited shorter average RTs.

DISCUSSION

Temporal error monitoring is a recent line of research at the intersection of timing, metacognition, and error monitoring. In a newly developed paradigm designed for rats[2] and reproduced in humans, we investigated commonalities and differences in temporal error monitoring between these two species. Both species consistently performed above chance level, but rats demonstrated superior choice accuracy in test trials, posing a question of which factors may underlie these discrepancies.

We hypothesized that better temporal error monitoring performance may correlate with a precise representation of task variables, reflected in their alignment. Humans and rats were expected to maintain representation of the "assigned" theta, a value determined by the experimenter for each human and rat based on the distribution of their temporal productions (TPs) in prior sessions, which governed the rule for reward assignment. The second variable, termed "observed" theta, was computed by applying the reward assignment rule to the TP distribution in the current session. Additionally, a third variable, the "choice" theta, was derived from the actual humans' and rats' selections between Small (SE) and (LE) ports/cards, considering the TP distribution in the current session. Indeed, we found that a smaller distance between the different thresholds correlated with higher choice accuracy. After matching data from rats and humans based on the distance between θ values, the interspecies discrepancy in choice accuracy disappeared and equalized at the session level. This indicates that the distance between different θ values plays a role in the higher performance for both species. Importantly, the correlation with choice accuracy was consistent across species, implying that both rats and humans represented the task structure similarly.

We also found that larger distances between TPs and theta values facilitated discriminability on a trial-by-trial basis, but with a significant difference in which task variable is guiding rats' and humans' choices. Humans based their decisions on "assigned" theta, whereas rats based their choices on "observed" theta. This suggests that rats make decisions based on the representation of the current distribution, whereas humans tend to rely on the representation of the threshold (rule) from the previous sessions. Previous studies, using varying paradigms, have also highlighted the ability of rats[19] and humans[20,21] to represent temporal distributions. Given the inherent difficulty of translational designs in subjecting humans and non-human species to identical regimes, it is possible that rats developed better representations of the full distribution of their responses, whereas humans, having undergone a smaller number of sessions, may have had less reliable representations of distribution and discrimination threshold.

We leveraged RTs to analyze decision-making in temporal error monitoring, as we assumed RTs may be an indicator of confidence judgments, also within the domain of temporal error monitoring.[2,18] Surprisingly, we found that RTs during the test trials reflected confidence only in rats. A larger distance from θ was associated with shorter RTs, suggesting that evidence accumulated faster as the current TP deviated further from θ. These results align with the lower choice accuracy observed in humans compared to rats. Several factors could explain this difference. One possibility is that if rats employ different timing mechanisms than humans, their error monitoring system may more easily decode these signals. A surprising consequence is that rats possess a more efficient temporal error monitoring system than humans. Yet another explanation is that decision-making strategies differ between species.

To test that hypothesis, we resorted to the comparison between species on training trials, where we found the same pattern between rats and humans. Both humans and rats responded faster to rewarded trials based on their anticipation of reward port matching with temporal error, albeit more prominently in rats, showing that they both had correct reward expectancies. This suggests that during training trials, rats and humans behave fairly similarly.

Differences between species became more pronounced when comparing RTs across training and test trials. Humans exhibited a larger RT increase from training to test trials, suggesting the engagement of additional processes during the test phase. This may account for the absence of a relationship between RT and choice accuracy in humans during test trials, in contrast to rats, where this association was preserved. One plausible explanation for the prolonged RTs in human test trials is species-specific differences in the tendency to explore versus exploit response strategies.[22] Additionally, task design could play a crucial role in shaping response patterns. In our study, humans and rats were required to choose between two discrete options, with their responses needing to align with a predefined rule to maximize points. This constraint may have led to more restricted RTs compared to the study by Kononowicz et al.[12] where a continuous-scale response format allowed humans and rats greater flexibility in expressing their confidence levels. Notably, in that study, the response format did not influence task outcomes, suggesting that the observed differences here in RTs may stem from the specific decision-making constraints imposed by the discrete-choice paradigm.

To assess confidence, we looked at RTs in the current study. However, an alternative way to assess confidence in temporal error monitoring responses was shown by Oztel and Balci,[13] who explored confidence and temporal error monitoring in

mice by measuring response rates (nose pokes). Following each timing response, they introduced a variable response window, with the assumption that nose poke rates during this window would reflect mice's confidence in their temporal accuracy. They found that response rates decreased with larger deviations from the target interval, following a negative quadratic trend, suggesting that mice were sensitive to both early and late errors. Oztel and Balcı[13] suggested nose poke rate as an alternative readout of confidence. It remains an open question for future research to assess differences between species in those alternative paradigms.

Despite species-level differences in performance, particularly during test trials, both rats and humans integrated temporal and reward information in a structured manner, demonstrating sensitivity to both task rules and reward contingencies. On a trial-by-trial basis, rats relied more on current time productions (TPs), while humans typically adhered to previously learned rules derived from earlier sessions. However, at the session level, both species exhibited a shared representation of the TP distribution, achieving comparable above-chance accuracy.

By using closely matched experimental designs in humans and rats, this study provides a direct comparison of behavioral strategies involved in temporal error monitoring. Although the present work does not aim to identify shared neural circuits or formal computational mechanisms, the cross-species behavioral analyses reveal consistent principles governing how both species evaluate their timing performance. This behavioral alignment establishes a common reference framework that can be leveraged in future studies combining large-scale neural recordings and computational modeling to investigate underlying mechanisms. Importantly, the identified similarities and differences between species highlight key constraints that must be considered when interpreting neural data. Beyond the specific domain of temporal error monitoring, these findings contribute to a broader understanding of error monitoring and self-evaluation in biological systems. They demonstrate that both humans and rats rely on structured internal representations of elapsed time, supporting the view that monitoring internal uncertainty is a fundamental and widely shared cognitive ability.

Limitations of the study

While the present study provides important insights into species differences in temporal error monitoring, several factors inherent to the comparison between rat and human experimental designs warrant consideration. One key difference is that humans completed substantially fewer sessions than rats, a factor that could potentially influence the stability of temporal representations, choice accuracy, and discrimination thresholds. In fact, rats need several sessions to be instructed in the use of the different elements of the task as well as the rules, a phase that was done using a different target duration and which conceptually matches the ''instruction'' phase in humans. When shifted to the new 2-s duration, rats also experienced more sessions as they progressively learned to apply the rule to the new duration, and supplementary analyses (see Figures S6A–S6C) demonstrate that the main species differences reported in Figure 1 remain when we consider the early sessions during testing. Nevertheless, these analyses reveal that experience modulates how task-relevant information is utilized across species (see Figures S6D and S6E). With increasing training, rats progressively rely more on immediate feedback to guide their choices, whereas humans rapidly integrate information from previous trials within a session to optimize decision-making. The observed species differences in temporal processing may thus reflect distinct learning strategies rather than simple disparities in training exposure.

Another difference is in the amount of muscle recruited to produce the target interval, which differs across species and between PRESS and HOLD behaviors. This difference is larger in rats than in humans (i.e., the whole body vs. hand) and may systematically influence timing performance. Rats are known to rely on motor routines for temporal decisions, and motor actions play a central role in time monitoring.[23–27] Evidence from both rodent and human studies further supports a link between motor processes and time perception.[18] In a previous human study,[18] we directly manipulated muscle involvement by comparing whole-arm joystick movements with simple button presses. Lower muscle engagement was associated with reduced motor variability and better temporal performance. These results suggest that greater muscle recruitment may introduce variability into timing signals, whereas reduced muscle involvement may support more precise temporal processing. In the original rat study, rats trained on a 3.2 s interval showed higher test-trial accuracy when time production involved lower active muscle mass (PRESS group). In the present study, these group differences were no longer observed when rats were trained on a shorter 2-s interval. This finding suggests that motor effects on timing depend on task duration and training history. It also indicates that the influence of motor involvement on temporal monitoring is not fixed. Further work is needed to clarify how sensorimotor factors shape timing and decision-making across species.

The present study also has limitations with respect to sex and gender. For the humans, recruitment was not stratified by sex as a primary criterion; rather, the near-equal distribution of men and women (16 men, 18 women) reflects the composition of the available volunteer pool, given the demanding multi-session nature of the study. As a result, the potential influence of sex or gender on temporal error-monitoring performance was not formally examined. For the animal model, only male Sprague-Dawley rats were included, consistent with the original study from which these animals were drawn. The absence of female animals, therefore, precludes any assessment of sex effects on timing behavior in the rodent sample. Future studies should incorporate sex-stratified designs in both human and animal cohorts to determine whether sex or gender moderates temporal error monitoring and the species differences reported here.

Finally, interspecies differences in motivational processing may have contributed to the observed behavioral disparities. Equating reward magnitude and subjective value across species remains a long-standing challenge in comparative cognition research. Motivational states may differ between animals and humans even under matched task conditions. Future studies should therefore examine how such motivational differences influence higher-order metacognitive processes, including temporal error monitoring.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Sena N. Bilgin (sbilgin@sd.psych.pan.pl).

Materials availability

This study did not generate new, unique reagents.

Data and code availability

- Data: All original datasets have been deposited at Zenodo and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- Code: All original code used for experimental tasks, statistical analyses, and figure generation has been deposited at Zenodo and is publicly available as of the date of publication. The DOI is listed in the key resources table.
- Additional information: Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This project has been funded by the National Center of Science in Poland (grant nr 2019/35/B/HS6/04389 to T.W.K.). We thank Bruce L. Brown for numerous enriching discussions about the study and his feedback on the previous version of the manuscript. The authors used Grammarly (an AI-assisted writing tool) for grammar and language correction during the preparation of this manuscript. The authors take full responsibility for the integrity and accuracy of all content.

AUTHOR CONTRIBUTIONS

S.N.B., T.W.K., and V.D. jointly conceived and designed the study. S.N.B. conducted the human experiments, while T.W.K. performed the animal experiments. S.N.B. and T.W.K. collaboratively carried out the statistical analyses, and V.D. contributed to the interpretation of the results. S.N.B. prepared the initial draft of the manuscript, and all authors provided critical feedback and contributed to the refinement of the research.

S.N.B. and T.W.K. contributed equally to this work and are recognized as co-first authors. Each played a central role in experimental design, data acquisition, analysis, and interpretation within their respective domains. Both assume joint responsibility for the integrity of the data, figures, and text, ensuring adherence to all editorial, ethical, and data reporting standards.

S.N.B., as the lead contact, coordinated data management, analysis, and manuscript preparation and will serve as the primary contact during review and after publication, including requests for materials, data, and clarifications.

All co-authors have read and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODELS AND STUDY PARTICIPANT DETAILS
  - Rats
  - Human participants
- METHOD DETAILS
  - Apparatus
  - Procedure
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Data preprocessing
  - Statistical modelling
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2026.115531.

Received: October 10, 2025
Revised: January 29, 2026
Accepted: March 26, 2026
Published: March 30, 2026

REFERENCES

1. Redish, A.D., Kepecs, A., Anderson, L.M., Calvin, O.L., Grissom, N.M., Haynos, A.F., Heilbronner, S.R., Herman, A.B., Jacob, S., Ma, S., et al. (2022). Computational validity: Using computation to translate behaviours across species. Philos. Trans. R. Soc. B *377*, 20200525. https://doi.org/10.1098/rstb.2020.0525.
2. Kononowicz, T.W., van Wassenhove, V., and Doyère, V. (2022). Rodents monitor their error in self-generated duration on a single-trial basis. Proc. Natl. Acad. Sci. USA *119*, e2108850119. https://doi.org/10.1073/pnas.2108850119.
3. Basile, B.M., Hampton, R.R., Suomi, S.J., and Murray, E.A. (2009). An assessment of memory awareness in tufted capuchin monkeys (*Cebus apella*). Anim. Cogn. *12*, 169–180. https://doi.org/10.1007/s10071-008-0180-1.
4. Beran, M.J., and Smith, J.D. (2011). Information seeking by rhesus monkeys (*Macaca mulatta*) and capuchin monkeys (*Cebus apella*). Cognition *120*, 90–105. https://doi.org/10.1016/j.cognition.2011.02.016.
5. Benjamin, A.S., Bjork, R.A., and Schwartz, B.L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. J. Exp. Psychol. Gen. *127*, 55–68. https://doi.org/10.1037/0096-3445.127.1.55.
6. Call, J. (2010). Do apes know that they could be wrong? Anim. Cogn. *13*, 689–700. https://doi.org/10.1007/s10071-010-0317-x.
7. Foote, A.L., and Crystal, J.D. (2007). Metacognition in the rat. Curr. Biol. *17*, 551–555. https://doi.org/10.1016/j.cub.2007.01.061.
8. Gherman, S., and Philiastides, M.G. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. eLife *7*, e38293. https://doi.org/10.7554/eLife.38293.
9. Rahnev, D., Koizumi, A., McCurdy, L.Y., D'Esposito, M., and Lau, H. (2015). Confidence leak in perceptual decision making. Psychol. Sci. *26*, 1664–1680. https://doi.org/10.1177/0956797615595037.
10. Jozefowiez, J., Staddon, J.E.R., and Cerutti, D.T. (2009). Metacognition in animals: How do we know that they know? Comp. Cogn. Behav. Rev. *4*, 29–39. https://doi.org/10.3819/ccbr.2009.40003.
11. Akdoğan, B., and Balcı, F. (2017). Are you early or late?: Temporal error monitoring. J. Exp. Psychol. Gen. *146*, 347–361. https://doi.org/10.1037/xge0000265.
12. Kononowicz, T.W., Roger, C., and van Wassenhove, V. (2019). Temporal metacognition as the decoding of self-generated brain dynamics. Cereb. Cortex *29*, 4366–4380. https://doi.org/10.1093/cercor/bhy318.
13. Oztel, T., and Balcı, F. (2024). Mice monitor their timing errors. Sci. Rep. *14*, 23356. https://doi.org/10.1038/s41598-024-71921-2.
14. Gallistel, C.R., and Gibbon, J. (2000). Time, rate, and conditioning. Psychol. Rev. *107*, 289–344. https://doi.org/10.1037/0033-295X.107.2.289.
15. Gibbon, J. (1992). Ubiquity of scalar timing with a Poisson clock. J. Math. Psychol. *36*, 283–293. https://doi.org/10.1016/0022-2496(92)90041-5.

16. Buhusi, C.V., and Meck, W.H. (2005). What makes us tick? Functional and neural mechanisms of interval timing. Nat. Rev. Neurosci. *6*, 755–765. https://doi.org/10.1038/nrn1764.

17. Shadlen, M.N., and Kiani, R. (2013). Decision-making as a window on cognition. Neuron *80*, 791–806. https://doi.org/10.1016/j.neuron.2013.10.047.

18. Bilgin, S.N., and Kononowicz, T.W. (2025). Temporal error monitoring: Monitoring of internal clock or just motor noise? Conscious. Cogn. *130*, 103849. https://doi.org/10.1016/j.concog.2025.103849.

19. Li, Y., and Dudman, J.T. (2013). Mice infer probabilistic models for timing. Proc. Natl. Acad. Sci. USA *110*, 17154–17159. https://doi.org/10.1073/pnas.1310666110.

20. McGuire, J.T., and Kable, J.W. (2012). Decision makers calibrate behavioral persistence on the basis of time-interval experience. Cognition *124*, 216–226. https://doi.org/10.1016/j.cognition.2012.03.008.

21. Jazayeri, M., and Shadlen, M.N. (2010). Temporal context calibrates interval timing. Nat. Neurosci. *13*, 1020–1026. https://doi.org/10.1038/nn.2590.

22. Laurie, V.J., Shourkeshti, A., Chen, C.S., Herman, A.B., Grissom, N.M., and Ebitz, R.B. (2024). Persistent decision-making in mice, monkeys, and humans. Preprint at bioRxiv. https://doi.org/10.1101/2024.05.07.592970.

23. Gouvêa, T.S., Monteiro, T., Soares, S., Atallah, B.V., and Paton, J.J. (2014). Striatal dynamics explain duration judgments. eLife *4*, e11386. https://doi.org/10.7554/eLife.11386.

24. Killeen, P.R., and Fetterman, J.G. (1988). A behavioral theory of timing. Psychol. Rev. *95*, 274–295. https://doi.org/10.1037/0033-295X.95.2.274.

25. Machado, A. (1997). Learning the temporal dynamics of behavior. Psychol. Rev. *104*, 241–265. https://doi.org/10.1037/0033-295X.104.2.241.

26. Meck, W.H., and Church, R.M. (1983). A mode control model of counting and timing processes. J. Exp. Psychol. Anim. Behav. Process. *9*, 320–334. https://doi.org/10.1037/0097-7403.9.3.320.

27. Merchant, H., Grahn, J., Trainor, L., Rohrmeier, M., and Fitch, W.T. (2015). Finding the beat: A neural perspective across humans and non-human primates. Philos. Trans. R. Soc. B *370*, 20140093. https://doi.org/10.1098/rstb.2014.0093.

28. Kilkenny, C., Browne, W.J., Cuthill, I.C., Emerson, M., and Altman, D.G. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. *8*, e1000412. https://doi.org/10.1371/journal.pbio.1000412.

29. World Medical Association (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. JAMA *310*, 2191–2194. https://doi.org/10.1001/jama.2013.281053.

30. Barr, D.J., Levy, R., Scheepers, C., and Tily, H.J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. J. Mem. Lang. *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

31. Gelman, A., and Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models (Cambridge University Press). https://doi.org/10.1017/CBO9780511790942.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
MATLAB	MathWorks	RRID: SCR_001622
Psychophysics Toolbox	OpenSource	RRID: SCR_002881
Graphic State 4	Coulbourn Instruments	RRID: SCR_012309
R software	R Core Team	RRID: SCR_001905
lme4 R package	CRAN	RRID: SCR_015654
Custom R scripts and datasets	User-developed	Zenodo Data and Code: https://doi.org/10.5281/zenodo.18940331
Other		
Rat strain: Sprague–Dawley	Envigo	RRID: RGD_734476

EXPERIMENTAL MODELS AND STUDY PARTICIPANT DETAILS

Rats

The animal study involved 16 male Sprague–Dawley rats (Envigo), aged seven months at the start of retraining to the 2-second target duration phase of the experiment. Only male animals were used in this study; the potential influence of sex on the results is acknowledged as a limitation in the Limitations section. The experiment was carried out in accordance with the guidelines established by the European Communities Council Directive (2010/63/EU) for compliance and use of laboratory animals and approved by the Paris-Saclay Institute of Neuroscience (NeuroPSI) institutional ethics board, the French Ministry of Research, and the French National Ethical Committee (reference: 2013/6). This study is reported in accordance with the ARRIVE (Animal Research: Reporting of *In Vivo* Experiments) guidelines[28] to ensure transparency and reproducibility in animal research.

All animals weighed 350–450 g at the start of the experiment, after which they were transferred to another protocol. The rats were housed in groups of four under a 12-hour light/dark cycle. After one week of acclimation with *ad libitum* access to food and water, they were food-restricted to maintain 85% of their free-fed weight. The animals were handled, weighed, and fed daily throughout the experiment. These rats were first trained with a 3.2-second target duration (data reported in Kononowicz et al.[2]). The current data (not previously published) are from the same rats, retrained with a 2-second target duration to enable direct comparison with humans trained under the same conditions. Two experimental sessions were conducted per day, separated by approximately three hours, during which the rats returned to their home cages with access to water.

Human participants

Thirty-four right-handed volunteers (16 men and 18 women, mean age = 26.89 years, SD = 6.50, range: 18–35 years) participated in the study. Race, ethnicity, and ancestry data were not collected, as these variables were not considered relevant to the research questions, and collection was not required by the approving ethics committee at the time of data acquisition. All participants self-reported normal or corrected-to-normal vision and hearing. Experimental procedures were approved by the Research Ethics Committee at the Institute of Psychology of the Polish Academy of Sciences (reference: IP.403.3.2022), and written informed consent was obtained from all participants before the experiment. All procedures were conducted in accordance with relevant guidelines, including the Declaration of Helsinki.[29] Participant recruitment was not stratified by sex as a primary criterion; due to the demanding multi-session nature of the study, recruitment was limited by practical availability, and an approximately balanced sample was achieved from the pool of eligible volunteers. The potential influence of sex on the study results was not formally analyzed and is acknowledged as a limitation (see Limitations section).

Each participant engaged in a temporal error-monitoring task presented in the form of a gamified paradigm, in which points earned during each session were converted to monetary rewards (0.3 PLN per point), provided via Virtual Sodexo Gift Cards. Participants were instructed to refrain from using external time-keeping strategies (e.g., tapping or counting) and were verbally briefed on the experiment and its reward structure.

METHOD DETAILS

Apparatus

Rats

For the rats, the experimental setup was built using the Habitest Modular System (Coulbourn Instruments) with an operant box placed in an isolation cubicle. Figure 1A, left panel, depicts the spatial arrangement of the behavioral box. On the left wall, a retractable lever

was placed between two reward ports, each connected to a feeder delivering grain-based pellets (45 mg, Phymep, Dustless Precision Pellets) and equipped with photocell sensors to measure port entries. The lever was separated from the ports by 7 cm and was positioned 6.5 cm above the floor grid. Reward ports contained a light-emitting diode to serve as an indicator for possible reward access after each successful TP production. A red house light and a speaker were placed on the right wall. The system was controlled, and data were collected using Graphical State 4 software, which provides a precision of 1 millisecond.

Human participants

Participants were seated 20 cm away from the monitor within the experimental room. The entire duration of the experiment involved presenting stimuli on a computer screen. The generation of experimental stimuli and the execution of tasks were controlled using Psychophysics Toolbox (Brainard, 1997) and MATLAB software. To respond to tasks, participants used the keyboard to press the spacebar, left arrow, or right arrow buttons.

Procedure

Overview of cross-species task structure

The temporal error-monitoring task originally developed for rats[2] was closely adapted for use with participants. In both species, the task followed the same conceptual workflow at each trial: the participant first produces a target time interval and then, immediately afterward, evaluates the magnitude of his/her/its own temporal error for the duration just produced at this trial (Figure 1A, see also Scheme S1). To achieve this stage, the overall structure of the experiment was nearly identical in rats and humans, and consisted of three phases: Duration Production Training, Temporal Error-Monitoring Training, and Temporal Error-Monitoring Test Sessions. For those seeking a comprehensive understanding of the rat pretraining process, we direct the reader to the details of the rats' pretraining and experimental procedures outlined in Kononowicz et al.[2] for further information.

Rat task

In the initial training,[2] the rats underwent a stepwise pretraining, starting with feeder training (two sessions) to familiarize them with the reward ports and the pairing of pellet drops with the onset of the reward port light. This was followed by port light training (ten sessions), where rats learned to nose poke in the lit port to trigger the delivery of a reward. Finally, lever press training (three sessions) required rats to press a lever to activate the reward cue in one of the ports, randomly chosen from trial to trial. These procedures gradually introduced the animals to the task. Rats then learned to produce a 3.2-s interval either by holding down a lever (HOLD group) or by executing two lever presses (PRESS group). At each trial, the lever was extended until the animal initiated its action and then retracted after the second action (releasing the lever for the HOLD group or pressing a second time for the PRESS group) and waited for an intertrial interval (ITI) of 4 to 8 seconds until the next trial (extension of the lever). Training progressed from very short intervals to the final target duration while imposing strict precision constraints: productions that were too short (TP < Target), or too long (TP + 2 $\times$ σTP of previous session) were unrewarded, and excessively long productions triggered an auditory penalty cue. This shaping procedure resulted in stable and precise temporal productions (TPs) (see Scheme S1A and S1D).

After reaching stability, the rats learned to monitor the magnitude of their temporal errors on a trial-by-trial basis. Each produced duration was classified as a small-error (TPSE) or large-error (TPLE) trial using individually computed thresholds, and these categories were associated with different reward magnitudes delivered respectively in one of the two ports (2 pellets vs. 1 pellet ports; sides counterbalanced between animals) (see Scheme S1B for details). On training trials, a given port was illuminated to indicate where the animal should poke to access the reward, but with a 50% chance of receiving it based on their just produced error (e.g., if the animal produced a TP within the range classified as TPSE, it will get 2p if the lit port corresponds to it, but 0p if the lit port corresponds to the TPLE). In dedicated test trials, both ports were illuminated, and the rat is free to choose to poke in either port and get rewarded if the chosen port corresponds to the accuracy of their own TP at that trial. Consistently selecting the high-value port after accurate productions (Figure 1C; TPSE-2p, green area) and the lower-value port after less accurate productions (Figure 1C; TPLE-1p, orange area) demonstrated that rats used internal error representations to guide behavior, indicating temporal error-monitoring abilities (see Scheme S1C for details). These initial data have been published[2] and are not reported here.

Building on the original paradigm, the same animals were retrained on a 2-s target interval. The retraining phase consisted of 4 sessions, followed by 23 sessions of Temporal Error-Monitoring Training. After each TP, the illuminated port indicated where a reward could be delivered (50%), following the same rule as the initial 3.2 s training. For each animal, an individual threshold (Θ; Figure 1C, black dashed line) was computed from its TP distribution to classify Small Error (SE; TPs between T and Θ) and Large Error (LE; TPs between Θ and 2 $\times$ σ(TP)) trials. SE trials were rewarded with 2 p in one port, and LE trials with 1 p in the other. Θ was set to divide SE and LE areas in a 1:2 ratio, based on the animal's performance in the last two precision-training sessions. Thresholds were updated if the proportion of SE to LE trials deviated from the optimal 0.5 ratio by a factor of 3 (i.e., below 0.16 or above 1.5) and recalculated if these bounds were exceeded in two consecutive sessions, ensuring balanced pellet delivery.

Following training, animals completed 34 Temporal Error-Monitoring Test sessions, each with 120 trials. Of these, 96 trials (80%) were training trials following the same probabilistic reward structure, while 24 trials (20%) were test trials in which both ports were illuminated, and reward delivery reflected the animal's temporal error evaluation. In test trials, rats could earn up to 48 additional pellets, for a total possible reward of 144 pellets per session. Choices between the two ports provided a direct measure of the animals' internal evaluation of temporal error. These data are reported here for the first time.

Human task

For humans, the experiment followed a similar structured sequence of phases over multiple sessions conducted across consecutive days. Each participant participated in 2 experimental sessions per day, each comprising 100 trials, with a total session duration of 30 minutes and a daily time commitment of about 1 hour. The procedure included 1 day of Duration Production Training, 2 days of Error Monitoring Training, and a minimum of 2 days of Error Monitoring Tests. In total, participants completed 2 sessions of duration production training, 4 sessions of error monitoring training, and at least 4 sessions of temporal error-monitoring tests (see Scheme S1D for details). Participants were expected to attend daily, though some experienced a 1-day delay due to personal circumstances. Before each temporal error-monitoring training and testing session, participants completed a pre-training phase, which involved 20 Duration Production Training trials to prepare them for the task. This pre-training was conducted before the start of each experimental day, except for the initial day.

Instructions and Reward Assignments

Participants were instructed to pay full attention to the experiment and not to use external time-keeping strategies such as tapping or counting. Participants were verbally briefed by the experimenter about the fundamental aspects of the experiment: (1) two experimental sessions would take place daily, encompassing 100 trials per session; (2) the session's duration would span between 20 to 30 minutes contingent on their performance; (3) after each session, a 60-second rest period would ensue, with the potential for extension based on their physical and mental state; (4) they also had the right to halt the experiment if they reported mental or physical fatigue. As the purpose of each stage of the experiment differed, participants were given different amounts of reward after each stage. Participants received a fixed reward amount during the Duration Production Training sessions, regardless of their performance. However, their rewards were contingent on their performance during the Temporal Error Monitoring Training and testing sessions.

Furthermore, participants were included in the experiment after all primary instructions were given. Throughout the experiment, task instructions were predominantly conveyed to participants through the computer screen, aiming to reduce direct interaction between participants and the experimenter. Nevertheless, the experimenter also guided participants based on their comprehension of the task instructions. This design choice aimed to maintain procedural similarity with the previous rodent study,[2] where animals had to infer task rules through outcome-based learning.

To investigate the extent to which the explicitness of task instructions affected participants' comprehension of the rule structure and their subsequent TP performance monitoring, we manipulated the clarity of task rules across two groups exclusively during the Temporal Error Monitoring (EM) Training phase, the only phase in which instructional content differed.

Error Monitoring Training Phase (No Instruction Group): Participants were instructed, ''If the time you produce is very close to the target time, you will earn 2 points, and if it is not so close to the target time, you will earn 1 point.''

Error Monitoring Training Phase (Instruction Group): Participant received a more explicit explanation of the rule mechanism: ''If the time you produce is close to the target time, you will get 2 points from one specific card displayed on the screen. If the time you produce is not very close to the target time, you will always get 1 point from the other card. You won't get any points in some trials regardless of your performance. However, an illuminated card still informs you about your performance, such that 0 points from the card where you earned a point means that it was a 1-point time production. 0 points from the card where you earned 1 point means that it was a 2-point time production.''

To sum up, the experiment investigated two main factors: variations in time production methods and the explicitness of task instructions. This resulted in the formation of four distinct human groups: PRESS/no-inst, PRESS/inst, HOLD/no-inst, and HOLD/inst. These groups allowed researchers to explore how different motor routines and instruction explicitness influence temporal error monitoring. Task instructions for each group were provided in the procedure section, denoted with quotation marks.

Target time presentation

In the pre-training session, participants were instructed to carefully observe and memorize the target interval (T), which was shown five times in a row and set at 2 seconds (T=2s). At the beginning of each representation, a white fixation cross appeared at the center of the screen, and then it turned green, indicating the start of the target interval. The on-screen time of the green fixation cross represented the interval. Then, the green fixation cross changed back to white color to prompt that the target interval had elapsed. The next trial was initiated after a random inter-trial interval (ITI) of 2 to 4 s.

Duration production training

During the Duration Production Training sessions, participants were instructed to reproduce the T interval presented in the pre-training phase. The level of instruction was not manipulated in this experimental phase, as the main goal was to train participants to learn how to reproduce time intervals. Each trial started with the presentation of the white fixation cross displayed in the middle of the screen. Participants were given unlimited time to start their time production (TP). Depending on the group to which the participants were assigned, they either had to generate TP by initiating the interval by pressing the ''Space'' key and terminating the interval by pressing the ''Space'' key again after they had sensed that the target time interval had already elapsed (Figure 1A right panel, PRESS group), or by pressing the ''Space'' key for the required time of target interval (HOLD group). Once participants started their TP, the white fixation cross on the screen was removed, and participants were shown a blank screen throughout their TPs. After a random inter-trial interval (ITI) of 2 to 4 seconds, if participants briefly pressed the keyboard in succession for 0.5 seconds or below, the trial was deemed invalid and restarted. We refer to such trials as invalid trials.

For each participant and session, the density function of the TP distribution was calculated, along with the ratio of the area under the curve exceeding the T criterion. Since a density function could not be computed for the first session of Duration Production Training, participants earned 1 point for TPs within the range (target, 2s) $\leq$ TP $\leq$ 4s, and 0 points for TPs outside this range. During the Training and Test Phases, the density function was calculated based on the TP distribution from the previous session. Participants' TPs that met the criteria 'T(2s) $\leq$ TP < 2σ(TP)s' (Figure 1C) were rewarded for both short and long productions. The purpose of this procedure was to motivate participants to be more precise in their TPs. Trials with TP outside of the criteria, 'TP < T(2s)' or 'TP $\geq$ 2σ(TP)s', were considered non-rewarded attempts, and the next trial restarted after a random 1.5s to 3.5s ITI. Participants were shown a 0.5s red fixation cross after long TPs, and "0p" after short TPs. On the other hand, after the rewarded trials ((2s) $\leq$ TP < 2σ(TP)), two identical cards were displayed on the left and right of the screen. One of these cards was randomly illuminated, and the participants were asked to collect their points by pressing the "right arrow" if the illuminated card was on the right, and the "left arrow" if it was on the left. Participants received 1 point for each trial, regardless of the accuracy of their time productions (TPs), unless the TP met the specified criterion. No time limit was set for the participants to collect their points. After participants collected their points, the next trial was initiated after a random 2 to 4s ITI. TPs below the target range (TP $\leq$ T, 2s), above the cut-off value (2* σ(TP) < TP), and invalid (TP < 0.5 s) were not included in the total number of trials (see Scheme S1A for more details). The purpose of this procedure was to provide equal training opportunities for each participant to produce a time close to the target range in a fixed design (two sessions of 200 trials).

Temporal Error Monitoring Training

The trial structure of the Temporal Error Monitoring Training phase was similar to the Duration Production Training phase up to the reward assignment part. At this stage, participants were trained to learn to assign a reward size (i.e., a reward assignment rule) as a function of TP error size in their TPs. Initiating from the training phase, deliberate manipulation of the instructional level led to the formation of two groups: the "no-inst" group and the "inst" group (see STAR Methods/instructions and reward assignments for details). To assign reward size to TP errors, TP density curves were calculated for each session. The threshold (θ, Assigned) (Figure 1C, dashed threshold line between green and orange areas) was estimated based on the participants' individual TP distribution in the previous session and assigned as the reward criterion for the subsequent session. θ was calculated by taking the integral of the area of the green and orange areas, dividing by 3, and then converting the value to the corresponding x-axis value. Figure 1C is a schematic representation of how rewards were assigned to specific parts of the TP distribution (also see Scheme S1B for more details). More accurate TPs between T and θ (we call them TP(SE) for TP Small Error) are rewarded with 2ps on one card (Figure 1C, green area). Less accurate TPs between θ and 2σ(TP) (we call them TP(LE) for TP Large Error) were awarded 1p on the other card (Figure 1C, orange area). The cards were randomly assigned to each participant as 1p (TP(LE)) and 2p (TP(SE)) cards and counterbalanced across participants. The locations of the cards were fixed for each participant across different phases. Generation of TP(SE) and TP(LE) on a given trial resulted in a 50% chance of receiving a reward as TP(SE)-2p or TP(SE)-0p and as TP(LE)-1p or TP(LE)-0p, respectively (see Scheme S1C for more details). The card-light cue did not always correctly indicate the reward card where a reward was available. Specifically, on a given trial, the card assigned as a 1p card delivered "1p" after a large error (Figure 1C, TP(LE)-1p, orange area) in rewarded trials but "0 p" after a small error (Figure 1C, TP(SE)-0-p, green area) in unrewarded trials, whereas the other card assigned as a 2p card delivered "2ps" after a small error (Figure 1C, TP(SE)-2p, green area) in rewarded trials but "0p" after a large error (Figure 1C, TP(LE)-0p, orange area) in unrewarded trials. After the participants chose the lit card by pressing the relevant keyboard key, the reward amount (points) was displayed by flipping the card on the screen, and the current trial was terminated (Figure 1C, training phase). The text "0 points" was not displayed on the screen in trials where the TP of the participants was less than T, and the trial restarted after a random ITI, to avoid confusion between non-reward trials and Duration Production Training trials where participants were not rewarded for poor TP performances (short and long TPs). Moreover, in trials where participants' TPs were longer than 2σ(TP), a red fixation cross was shown, and the trial was restarted after a random ITI. During the training phase, participants were expected to earn between 50 and 100 points, depending on their performance in temporal productions (TPs). If participants produced small error TPs (2p reward), they could earn a maximum of 100 points per session. In contrast, if participants produced larger error TPs (1p reward), they would earn fewer points, with a minimum of 50 points per session, reflecting the standard reward allocation for normal productions. This point system was designed to motivate participants to produce more accurate TPs while ensuring a consistent and fair reward structure across sessions, depending on their performance in TPs.

Temporal Error Monitoring Test

To test whether humans can report their temporal errors, we used the same paradigm as in the Temporal Error Monitoring Training phase in 70% of trials, and in the remaining 30% of trials, we presented test trials in which both cards were illuminated (Figure 1C, test phase; Scheme S1C). For the rats, 80% of trials followed the same structure as in the Temporal Error Monitoring Training phase, while the remaining 20% were used to assess their capacity to report temporal errors. The main purpose of the test trials was to allow participants to choose one of the cards/ports according to the reward-assignment rule described above. Participants were presented with the opportunity to receive rewards from both cards/ports with different reward sizes (0 vs. 1 vs. 2 ps (points for humans/pellets for rats)) depending on their TPs and according to the rule. No time limit was set for participants during the choice phase.

To maximize the reward rate, participants had to select cards/ports considering the accuracy of their time productions (TP), classified as TP(SE) or TP(LE), and the card/port position associated with the corresponding reward magnitude. Briefly, choosing the "2 ps" (Small Error, SE) card/port in TP(SE) trials or the "1 p"(Large Error, LE) card/port in TP(LE) trials was considered a demonstration of the participants' error monitoring abilities.

In each training session, participants had the potential to earn up to 70 points (as 70% of trials were training trials). These points were awarded for producing TPs within the designated criteria. For instance, the participant could earn a maximum of 2 points per TP trial that met the accuracy criteria, with a 50% chance of receiving the reward. In contrast, test trials differed from training trials in that every TP meeting the specified criterion was rewarded. In addition to training points, participants received rewards as long as they selected the correct card on test trials; therefore, participants could earn up to 130 points in total during one session of the testing phase.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data preprocessing

All analyses were performed using custom-written R scripts (version 2023.12.1+402) and were applied consistently to both rat and human data.

Similar to the previous study,[2] TPs below 1.25 s were excluded to eliminate impulsive responses and motor errors, ensuring a more accurate assessment of performance. Z-score normalization was applied to TPs and RTs for each human and rat individually. For each human and rat, the mean and standard deviation of TPs and RTs were calculated within each session. Z-scores were then computed by subtracting the mean of each variable from the raw score and dividing by the standard deviation. This process was applied to each human and rat separately within each session, ensuring that individual differences in TPs and RTs were normalized for comparison. For TPs, z-scores exceeding 3 at both the session and human and rat levels were excluded. For RTs, only z-scores exceeding 3 at the session level were removed, as variations in RT profiles among humans and rats were considered natural. Outliers were further removed by excluding RTs exceeding 10 s.

For the analyses reported in Figure 3, RTs were further normalized to account for individual differences in baseline speed. The mean RT was calculated separately for left- and right-side responses for each rat and human. Each trial's RT was then centered by subtracting each rat's and human's side-specific mean, producing a normalized RT value that reflects deviations from the individuals' typical response for that side. This approach controls for overall speed differences between humans and rats and accounts for potential biases between left and right responses. Using these normalized RTs allows for more accurate comparisons across conditions, rewarded versus test versus no-test trials (Figure 3A) and unrewarded trials (Figure 3B).

Sessions in which "2p" or "1p" port/card choices accounted for less than 16.6% of the total were considered imbalanced and excluded from analysis. This threshold corresponds to a minimum of 5 "2p" or "1p" choices out of 30 test trials, ensuring sufficient representation of both options. Two participants (participants 8 and 32) were entirely excluded due to a consistent global preference for the "1p" option, indicating a strong selection bias.

Initially, 188 testing sessions were conducted with humans, each completing at least four sessions (see the procedure section for details). After applying the exclusion criteria, 115 sessions remained. In the rat dataset, 544 testing sessions (each lasting 2 seconds) were conducted, with an average of approximately 34 sessions per rat (16 rats in total). Following the same exclusion procedures, 330 sessions were retained. In total, the final dataset included 50 participants and 445 valid test sessions for analysis.

For the analyses, data from groups exposed to two different time-production and instruction methods were merged. Analyses confirmed no significant differences in TP performance or choice accuracy across groups (see Figures S4 and S5), justifying pooling the data within each species.

Statistical modelling

To account for multiple observations per human and rats across various sessions, Linear Mixed-Effects Models (LMMs) estimated through Restricted Maximum Likelihood (REML) with the *nloptwrap* optimizer were employed. Additionally, Generalized Linear Mixed-Effects Models (GLMMs) were used via the ***lme4*** package (version 1.1-21).

LMMs are regression models that account for fixed and random effects for continuous response variables (e.g., TPs and RTs). Fixed effects are assumed to be consistent across experiments, while random effects represent random variation within the population.[30] GLMMs are regression models for binary response variables (e.g., port/card choice, choice accuracy).[31]

The mixed-effects modeling approach was combined with model comparison, allowing for the systematic selection of the most appropriate model. These comparisons were conducted using the Akaike Information Criterion (AIC), which identifies the model that best explains the variability in the data. In addition, improvements in model fit were formally assessed using likelihood ratio tests via ANOVA. In line with the methodology employed in the earlier study by Kononowicz et al.,[2] where multiple experimental sessions were conducted per human and rat, GLMMs and LMMs incorporated a nested random effect structure, accounting for potential variations in each human's and rat's performance across different sessions.

For single-trial level analyses, a model was established for each human and rat that included only the intercept term. Various random effects structures were compared using AIC. As a result, "Human, or Rat" and "Sessions" were incorporated as random effects in the single-trial level analyses. In contrast, single-session level analyses included only "Human, or Rat" as random effects. In all models, standardized parameters were obtained by fitting the model on a standardized version of the dataset, and P-values were computed using a Wald t-distribution approximation.

In all figures, asterisks denote significance levels as follows: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$. Non-significant results are indicated as 'ns'.

Time production analysis (related to Figure 1): The mean (μ) and standard deviation (σ) of TPs were computed across sessions and humans/rats. To compare TP performance (mean TPs, μ(TPs)) across humans/rats. and sessions, an LMM was employed. The model predicted μ(TPs) with "Species" as a fixed effect (formula: "μ(TPs) ~ Species"). To investigate timing precision (standard deviation of TPs, σ(TPs)), a parallel LMM was applied, predicting σ(TPs) with "Species" as a fixed effect (formula: "σ(TPs) ~ Species"). Both models included "Human, or Rat" as a random effect (1|Human, or Rat) to account for individual variability. This analytical approach facilitated a comprehensive investigation into the mean and precision of temporal judgments across different species while accounting for individual variations. By normalizing TPs and incorporating both fixed and random effects, we ensured that our analysis robustly compared species-specific differences in time production performance and precision.

Current TP & TP history analyses (related to Figure 1): In our analysis, we utilized z-score criteria to evaluate TPs for each human/rat and session, enabling a direct comparison of performance across different species. We calculated the current-trial TP (denoted as TPn) and the TP history (represented by the mean TP of the past 10 trials, μTP10) based on these z-scores. Subsequently, we fitted a GLMM to predict the probability of selecting a small error (2ps) port or card. The predictors included the current-trial TP (TPn), TP history (μTP10), and species, with the interaction terms to explore potential species-specific effects (formula: "Choosing 2p Port/Card ~ TPn * Species + μTP10 * Species"). "Human, or Rat" and "Sessions" were incorporated as random effects to account for within-human/rat variability and potential session-specific influences. The addition of the species factor to the model significantly improved its explanatory power, as evidenced by a substantial reduction in the Akaike Information Criterion (ΔAIC = 171, **p = .001). This indicates that the inclusion of species provided a better fit to the data. Incorporating "Sessions" as random effects further accounted for potential variability and dependencies within the data, enhancing the model's robustness.

Choice accuracy on test trials (related to Figure 1): An LMM was applied to predict choice accuracy values based on species, using the formula: "Choice Accuracy ~ Species". "Human, or Rat" were included in the model as a random effect to account for individual variations. This analytical approach provided a robust framework for assessing the influence of species on session accuracy while accounting for the variability among individual humans and rats.

Session-level task variables (related to Figure 2): Task variables at the session level play a crucial role in enhancing error monitoring capabilities, contingent upon the ability to comprehend diverse aspects of these variables and establish a mental representation of the designated reward rule. To facilitate a comparison between species, theta values (θ) were normalized by dividing by the overall precision of each rat and human, indicated as σ(TPs) across sessions. This normalization allowed for the comparison of task representations among individuals, regardless of their precision levels.

Specifically, the outcomes of operations on the thetas were divided by the overall precision of each rat and human, σ(TPs), across sessions to ensure an unbiased comparison of task representations among individuals. Additional adjustments were made by adding 1 to the results to prevent undefined values. Following these computations, z-score criteria were applied to each theta operation, excluding z-scores exceeding 3 to eliminate outliers. Finally, a log transformation was applied to normalize the obtained data.

Single-session-level computations: The normalization process was conducted using the following formulas to standardize task variables relative to each human's and rat's overall precision (σ(TPs)) across sessions, as shown in the Alignment Index Equations 1, 2, and 3. This approach facilitated unbiased comparisons among individuals regardless of their precision levels. The steps involved were:

1) Normalization and Adjustment: For each task variable, we first computed the absolute difference between pairs of thetas, normalized it by the overall precision (σ(TPs)) of the humans and rats, and then added 1 to avoid undefined values.
2) Log Transformation: After normalization, we applied a log transformation to further standardize the data.

The specific formulas used were:

$$Alignment\ Index\ 1 = log\left(\frac{|\,Assigned\ \theta - Choice\ \theta\,|}{\sigma(TPs)_{subject}} + 1\right) \quad \text{(Equation 1)}$$

$$Alignment\ Index\ 2 = log\left(\frac{|\,Assigned\ \theta - Observed\ \theta\,|}{\sigma(TPs)_{subject}} + 1\right) \quad \text{(Equation 2)}$$

$$Alignment\ Index\ 3 = log\left(\frac{|\,Observed\ \theta - Choice\ \theta\,|}{\sigma(TPs)_{subject}} + 1\right) \quad \text{(Equation 3)}$$

3) Outlier Removal: After these computations, we applied z-score criteria to each theta operation, excluding z-scores exceeding 3 to eliminate outliers, ensuring a robust and normalized comparison of task representations among individuals.

Session Level Matching Procedure: To facilitate precise cross-species comparisons, we implemented a thorough matching process for the theta alignment index data in the human group, aligning each data point with its closest rat counterparts in every iteration. This iterative approach guaranteed the uniqueness of session selections in each iteration, preventing their reuse and maximizing the

data collected from the rat group. In cases where multiple rat sessions were equidistant from a given theta index in the human data, we adopted a random selection approach to maintain an unbiased process.

After initial exclusions, the dataset consisted of 445 sessions in total (330 from rats and 115 from humans) (see Data Processing section for details). To improve data quality, we removed sessions in which the theta alignment index was considered an outlier (i.e., more than 3 standard deviations from the mean). This outlier removal resulted in a final set of 327 rat sessions and 91 human sessions. To enable balanced comparisons across species and experimental conditions, a matching procedure was then applied. This reduced the dataset to 364 sessions, with 91 sessions assigned to each of the four resulting groups: Human–Assigned Choice, Human–Observed Choice, Rat–Assigned Choice, and Rat–Observed Choice. The resulting design was fully balanced, with an equal number of sessions across species (human, rat) and task conditions (Assigned, Observed), allowing for fair and controlled analyses.

To evaluate the consistency of choice accuracies based on the theta alignment indices across different species, we employed an LMM. The model incorporated the formula: ''Choice Accuracy $\sim$ Theta Alignment * Species'' with ''Human, or Rat'' as a random effect (1 | Human, or Rat) to evaluate the consistency of choice accuracies based on these indices across species.

Moreover, to explore the relationship between TP performance monitoring in a single trial and how humans and rats aligned their theta indices (see Figure S3), we conducted an LMM analysis. This analysis employed the current trial TP (TPn) and the previous 10 TPs (μTP10) as predictors for the probability of selecting the 2p (Small Error, SE) port or card. The analysis was performed separately for each human/rat and each session individually (no random effects in this case) using the formula: ''Choosing 2p Port/Card $\sim$ TPn + μTP10''. To refine the data, we excluded sessions where ''1p'' or ''2p'' choices accounted for less than 20% or more than 80% of total choices. Subsequently, z-score criteria were applied to the estimated values of the current-trial TP and TP history within each species, removing values exceeding a z-score of 3 from the analysis. The estimated values derived from this LMM analysis served as predictors in a GLMM, with the formula: ''Theta Alignment$\sim$TP Estimates * Species''. ''Human, or Rat'', were included as random effects in this comprehensive analytical approach to account for individual variability. This model aimed to evaluate how TP performance monitoring influences theta index alignment across species.

Single-trial-level computations: On an individual trial basis, task variables were determined by assessing the distance between three distinct theta values and the TP of the current trial, as shown in the Distance Index Equations 4, 5, and 6. Acknowledging the potential proximity and overlap of these theta values, an exploration was conducted to understand their impact on task representation for each trial.

The three Distance Indices were defined as follows:

$$Distance\ Index\ 1 = TP_{current} - Assigned\ \theta \quad \text{(Equation 4)}$$

$$Distance\ Index\ 2 = TP_{current} - Observed\ \theta \quad \text{(Equation 5)}$$

$$Distance\ Index\ 3 = TP_{current} - Choice\ \theta \quad \text{(Equation 6)}$$

To investigate the influence of these task variables on the probabilities of choosing the correct port or card on a single trial, GLMMs were employed. Each task variable (''Assigned'', ''Observed'', and ''Choice'') served as a predictor for the probabilities, with species specified as a factor (formula: ''Choice Accuracy $\sim$ Theta Distance * Species''). The model accounted for random effects by incorporating ''Human, or Rat'' and ''Sessions''. This analytical approach provided a nuanced understanding of how the defined task variables influenced the decision-making process across species.

RTs as an index of confidence (related to Figure 3): For session-level analysis, a GLMM was employed, incorporating RTs as a predictor for choice accuracy values in two distinct species. The formula used for this model was ''Choice Accuracy $\sim$ μ(RTs) * Species''. ''Human, or Rat'' were included as random effects.

For trial-level analysis, a GLMM was applied to model RTs as a predictor for selecting the correct port/card on a per-trial basis. The formula for this model was ''Choice Accuracy $\sim$ RTs * Species.'' In this case, both ''Sessions'' and ''Human, or Rat'' were considered as random effects. This dual-level analysis provided a comprehensive exploration of the relationship between Response Times and choice accuracy at both session and trial levels, accounting for individual variations within humans and rats and sessions.

ADDITIONAL RESOURCES

No additional resources apply to this study.

5.2) Paper 2 : *Monitoring of Internal Clock or Just Motor Noise?*

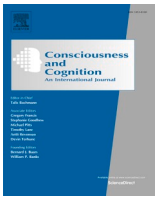
Consciousness and Cognition 130 (2025) 103849



Contents lists available at ScienceDirect

Consciousness and Cognition

journal homepage: www.elsevier.com/locate/yccog



Full Length Article



Temporal error monitoring: Monitoring of internal clock or just motor noise?

Sena N. Bilgin [a], Tadeusz W. Kononowicz [a,b,*]

[a] Institute of Psychology, Polish Academy of Sciences, Ul. Jaracza 1, 00-378, Poland
[b] Université Paris-Saclay, CNRS, Institut des Neurosciences Paris-Saclay (NeuroPSI), 91400 Saclay, France

ARTICLE INFO

Keywords:
Error monitoring
Metacognition
Interval timing
Motor timing
Time production

ABSTRACT

Understanding how humans monitor and evaluate temporal errors is crucial for uncovering the mechanisms of metacognitive processes, linking the fields of time perception and metacognition. In a typical paradigm, participants self-generate a time interval and subsequently can accurately evaluate its error. The implicit assumption in the field has been that participants monitor temporal representations. Even though temporal error monitoring has been replicated numerous times, it remains unclear what kind of information participants monitor when assessing the just-generated interval. Here, we assessed two scenarios in which participants could monitor sources of variability in temporal error monitoring: the internal representation of duration (Clock Hypothesis) or just motor signal (Motor Hypothesis). We assessed temporal error monitoring by inducing different levels of motor signal in motor timing, with the expectation that these levels of motor execution would influence temporal error monitoring outcomes. The motor signal was manipulated by instructing participants to either use button presses or joystick movements to produce time intervals, allowing us to evaluate and report how different levels of motor execution signal affect temporal error monitoring. According to the Clock Hypothesis, the additional motor signal should impair the accuracy of temporal error monitoring. Conversely, the Motor Hypothesis posits that additional induced signal should enhance the accuracy of temporal error monitoring. In line with the Clock Hypothesis, error monitoring performance was enhanced in a condition with a lower motor signal. These results show that humans evaluate their errors based on an informationally rich representation of internal duration, supporting metacognitive abilities in temporal error monitoring.

Public significance: Temporal error monitoring emerged from the fields of interval timing, decision-making, and metacognition, positing that humans evaluate the sign and magnitude of their temporal errors. Here, we critically test whether participants assess their timing representations as such and whether they are aware of the correctness of these evaluations.

1. Introduction

Metacognitive monitoring is often defined as the ability to accurately evaluate one's cognitive performance without explicit feedback (Yeung & Summerfield, 2012). A broad spectrum of information can be subject to metacognitive monitoring, such as

* Corresponding author at: Institut des Neurosciences Paris-Saclay - NeuroPSI CNRS, Centre CEA Paris Saclay, RD306 entrée Sud, Allée des Neurosciences, bâtiment 151, 91191 Gif-sur-Yvette, France.
E-mail address: tadeusz.kononowicz@cnrs.fr (T.W. Kononowicz).

https://doi.org/10.1016/j.concog.2025.103849
Received 1 November 2024; Received in revised form 20 January 2025; Accepted 11 March 2025
Available online 24 March 2025
1053-8100/© 2025 Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

perceptual information from the external world (e.g., visual, auditory) or internal bodily signals (e.g., interoception). The sense of time is an example of such an internal signal (Paton & Buonomano, 2018). However, metacognitive monitoring of errors in our timing behavior has only recently been discovered (Akdoğan and Balcı, 2017; Kononowicz et al., 2019), constituting a promising avenue of research into how the brain monitors internally generated temporal information.

Tracking time intervals is crucial for many activities, from predicting when a traffic light will turn red to making split-second decisions in sports. In the time production task, humans are asked to generate previously learned time intervals by pressing a response button for a specified amount of time or interleaving a specified amount of time between two presses (Getty, 1976; Macar et al., 1999). While the timing of these button presses is accurate on average, internal uncertainty in timing representations leads to significant variability in errors from one trial to another (Balci et al., 2011; Jazayeri & Shadlen, 2010; Mamassian, 2008). The discovery of temporal error monitoring has shown that humans can introspect about their timing errors on single trials by reliably indicating the direction or magnitude of their error after each production (Akdoğan & Balcı, 2017; Kononowicz et al., 2019, 2022). More specifically, after each time production, participants report the direction (too short, too long) and magnitude of their timing errors on a continuous scale (Kononowicz et al., 2019). The estimated magnitude of timing errors correlates with time production, suggesting that the error in their production is accessible. These results lead to claims about the metacognitive nature of temporal representations in humans and the ability to introspect about temporal errors (Kononowicz et al., 2019; Oztel & Balci, 2023). Despite numerous publications replicating the effect in various forms (e.g., Bader & Wiener, 2021, 2024, Duyan & Balcı, 2019, 2020; Yallak & Balcı, 2021; see also Öztel et al., 2021; Öztel & Balcı, 2023), a fundamental question persists: what kind of information do subjects evaluate when they engage in self-evaluation of temporal errors (also see Jovanovic et al., 2021; Pavailler et al., 2024)?

The prominent Wing & Kristofferson (1973) model proposes that variance in motor timing tasks can be decoupled into motor and clock variance. It posits that consistent time intervals rely on two primary components: a time-keeping process and a motor implementation process, which is typical for time-production tasks. In line with this view, we formulated two hypotheses (although not mutually exclusive) in a temporal error monitoring task: participants can monitor signals related to the internal clock (Clock Hypothesis) or the motor signal related to motor execution (Motor Hypothesis). The two hypotheses are not mutually exclusive. They rather represent two ends of the continuum. It is plausible that participants read out both signals in error monitoring. We hypothesized that manipulating motor signal variance would alter the relationship between timing and motor signals. To test this, we experimentally varied the motor signal component to assess temporal error monitoring performance across Press and Move conditions.

According to the Clock Hypothesis, the additional induced motor signal in the Move condition should impair the accuracy of temporal error monitoring. Conversely, the Motor Hypothesis posits that additional induced signal in the Move condition should enhance the accuracy of temporal error monitoring. To test whether these assumptions generalize to shorter time intervals and different sensorimotor demands, we asked participants to judge the length of their reaction times (Corallo et al., 2008; Marti et al., 2010; Bratzke & Bryce, 2016). Participants engaged in both tasks using either their dominant or non-dominant hand. We hypothesized that the predictions of the Clock and Motor hypotheses would generalize across both Time Production (TP) and Reaction Time (RT) tasks. As participants reported their confidence in their error judgments, we predicted that their error judgments would be more accurate at high levels of confidence, suggesting hierarchical introspective access to temporal judgments and the accuracy of these judgments.

Note that predictions regarding the main experimental manipulations primarily focus on temporal error monitoring. While research indicates that motor movements can improve timing performance, other studies reveal inconsistent biases of various movement modalities on interval timing (De Kock et al., 2021). Consequently, we did not establish a clear prediction regarding the influence of motor movements on-time production.

2. Methods

2.1. Participants

The experiment involved thirty-one volunteers, primarily right-handed (19 females, $M_{age} = 25.70$, $SD_{age} = 5.74$ years), with normal or corrected-to-normal vision and hearing. Written consent was obtained from all participants, and approval was granted by the Research Ethics Committee at the Institute of Psychology of the Polish Academy of Sciences (Nr: IP.403.3.2022). Participants received 100 PLN as compensation via Virtual Sodexo Gift Cards.

Participants completed multiple blocks in a single day, including a pretraining block, two training blocks (one focused on practicing the main tasks—time production (TP) and reaction time (RT)—and the other on temporal error monitoring and confidence rating) and three experimental blocks (two dedicated to time production and one to reaction time). Each experimental block consisted of 80 trials, while each training block included 12 trials. The total duration of the experiment, including training, was approximately 1.5 h.

2.2. Apparatus

Participants were positioned 45 cm from the screen in a relaxed sitting position with their elbows on the desk to keep the equipment stable. If they sat straight and moved closer to the screen, they were positioned 30 cm away. The creation of experimental stimuli and task execution were facilitated using Psychophysics Toolbox (Brainard & Vision, 1997) and MATLAB software. Throughout the experiment, participants used a Thrustmaster Sidestick Airbus Edition joystick to respond to the task. During the “Move” trials, performed a motor action by moving the entire joystick downward by 7.5 cm from its center position, requiring coordinated hand and arm movement. In contrast, during “Press” trials, participants executed a fine motor action by pressing the front button on the joystick,

which involved a finger press requiring a displacement of approximately 1 cm to register the response. These different motor demands were designed to elicit varying levels of motor execution signal. Data were then converted to text files and analyzed using custom-written R code.

2.3. Procedure

Before the experiment began, participants received clear instructions emphasizing the importance of giving their full attention to the tasks and avoiding any attempts to count time during the time perception experiment. They were informed that their performance could vary based on daily mental and physical conditions and might not necessarily reflect their true abilities, but their data would still contribute to scientific research. The experimenter verbally outlined the overall duration of the block and the estimated length of each block. Additionally, both the experimenter and the computer screen provided task-specific instructions for each experimental phase and block, ensuring that participants understood the requirements throughout the block.

The main experiment investigated how variations in motor actions influence temporal error monitoring by manipulating motor demands and handedness (Wang et al., 2018). Specifically, motor demands were varied with Press trials (low signal) and Move trials (high signal), while handedness was manipulated with DH trials (low signal) and NDH trials (high signal). DH trials involved motor actions performed with the dominant hand, typically used for most activities, whereas NDH trials involved the non-dominant hand. Participants were randomly assigned to experimental conditions within a single day, using a 2x2 within-subjects design.

2.4. Pre-training phase

During the pre-training block, participants were shown a target interval of 1.5 s (T) displayed consecutively 10 times. Each trial began with a white fixation cross (2 cm in size) displayed in the center of the screen. The cross then changed to green to signal the start of the target interval, which lasted for the full 1.5 s. After the target interval, the cross turned back to white to indicate the end of the interval. Subsequent presentations were separated by an inter-trial interval (ITI) ranging from 1 to 3 s.

2.4.1. TP Task Training Phase

The primary goal of the duration production training phase was to instruct participants to accurately reproduce the target time interval. Each trial began with the presentation of a cue: relevant text indicating the assigned motor action (“Press” or “Move”) positioned on either the left or right side of the screen, depending on the handedness condition. After a random delay of 1–3 s (inter-stimulus interval, ISI), participants either pressed a button on the joystick or moved the joystick. In the Press condition, they defined the time interval by performing two consecutive button presses: the interval began with the onset of the first press and ended with the onset of the second press. “Onset” refers to when the time when the button was pressed, not released. In the Move condition, participants initiated the time interval by moving the joystick downward from the center position and then returning it to the center. A back-to-center movement defined the start of a timed interval. To mark the end of the interval, participants performed the same sequence. Participants were instructed to move the joystick only downward, with any deviation resulting in trial restarts. Participants adhered to instructions to perform the Press or Move conditions with the corresponding hand according to the left/right side position of the given cue. Participants encountered four conditions during the experiment: Press/Dominant-Hand, Press/Non-Dominant-Hand, Move/Dominant-Hand, and Move/Non-Dominant-Hand. Each block included all four conditions, with the order of conditions randomized within each block to control for potential order effects.

Once participants initiated their target precision (TP) tasks, the white fixation cross disappeared, and a blank dark gray screen (RGB: 169, 169, 169) was displayed for the duration of the TP. Participants were instructed to end their TPs when they perceived the target time had passed. Trials where participants produced a duration below 0.5 s were considered invalid and restarted. For each participant, the mean and standard deviation of TPs from the current block were calculated to analyze performance and determine feedback ranges. During training, since an upper limit specific to each participant couldn’t be determined, a cutoff value of 3 s was set for TPs. TPs exceeding this limit were disregarded, and the trial was restarted. Participants completed 12 trials during the training phase, with 3 trials for each of the four conditions. The trials were mixed within the training phase to ensure that participants experienced all conditions in a randomized order.

2.4.2. RT Task Training Phase

The primary goal of the reaction time (RT) training was to ensure that participants understood the task and responded correctly to the stimulus. Each trial began with an informative cue: text indicating the required motor action, displayed randomly on the left or right side of the screen to indicate handedness. Following a random inter-stimulus interval (ISI) of 1–3 s, allowing time to process the cue, a 3.5 cm white square appeared at the bottom center of the screen, prompting an immediate reaction. In the Press trials, participants were instructed to press a button as soon as the square appeared. However, to prevent premature responses, participants were explicitly instructed not to press the button continuously before the trial started or during the ISI. Trials were terminated if these rules were violated, which could have contributed to the slightly delayed RTs as participants ensured they were responding only after the visual cue appeared.

In the Move trials, participants were required to perform a downward joystick movement toward the square. To ensure consistent performance, they had to maintain the same vertical trajectory to reach the target, eliminating deviations such as curved movements. Any deviation from this required trajectory resulted in the trial being eliminated and restarted. Also, trials were terminated if reaction times exceeded 4 s. To evaluate performance, each participant’s RT’s mean, and standard deviation were calculated during this phase

to establish feedback criteria for subsequent experiment stages. Participants completed 12 mixed trials, with 3 trials for each condition, ensuring that all conditions were represented and randomized within the training phase.

2.4.3. Error Judgement Training Phase

The primary goal of the Error Judgement Training Phase was to familiarize participants with the procedure for reporting their TP and RT errors on a single-trial basis. During the TP task, participants received immediate feedback after completing their TPs and the subsequent brief action. This feedback was presented as text in the center of the computer screen and used numerical values along with "-" and "+" signs to indicate the magnitude of the error and whether the productions were shorter or longer compared to the target time. TPs falling between 2 and 3 standard deviations from their means were rated as "-0" or "+0," indicating productions that were "too short" or "too long." TPs falling between 1 and 2 standard deviations were rated as "-1" or "+1," reflecting "considerably shorter" or "considerably longer" productions. TPs falling between 0.5 and 1 standard deviation were rated as "-2" or "+2," indicating "shorter" or "longer" productions, while TPs falling between 0 and 0.5 standard deviations were rated as "-3" or "+3," indicating "slightly shorter" or "slightly longer" productions. When participants accurately produced the target interval of 1.5 s, the display showed "3*," indicating an "exact" production. TPs deviating more than 3 standard deviations from the mean resulted in trial restarts as a penalty.

In the RT task, the training procedure paralleled the TP task, with feedback based on participants' RTs. However, since participants could not react before the onset of the visual stimulus, feedback was given based on the speed of their reactions to the visual stimulus onset. The same rating scale procedures were applied, with "+3", "+2", "+1", and "+0" indicating "very quick", "quick", "considerably slow", and "too slow" reactions, respectively. RTs deviating more than 3 standard deviations from the respective means also led to trial restarts.

Similar to the previous training phases, participants were expected to complete 12 trials for both tasks, evenly distributed across the conditions with 3 trials per condition.

2.4.4. Error judgement test phase

During the test phase, we aimed to assess participants' ability to report temporal errors in their TPs and RTs, as well as their confidence in these reports. Previous studies have used separate categorical response scales to measure both confidence (based on overall performance accuracy) and error monitoring performance (e.g., Akdoğan & Balcı, 2017; Yallak & Balcı, 2021). We utilized a combined response scale to measure confidence in error monitoring, along with assessing error magnitude and directionality on a continuous scale (e.g., Kononowicz et al., 2019; Öztel & Balcı, 2022). After participants completed their time production with the second brief motor action, they were presented with a new coordinate system-based Error and Confidence Judgments Rating Scale. The x-axis is represented as a linear scale from "-" (shortest duration) to "+" (longest duration), indicating the temporal position of the reproduced time relative to the target time. The y-axis ranged from 0 % to 100 %, representing explicit confidence in monitoring decisions. Participants were instructed to make these two judgments and mark the judgment position by moving the joystick toward the coordinate intersection and then pressing the front button of the joystick. Following participants' decisions, feedback on their monitoring performances was provided in 50 % of the trials. The next trial started within a random ITI ranging from 1 to 3 s.

In the control experiment, the RT task served as a baseline condition for exploring temporal error monitoring during self-initiated motor actions. As in the TP task, participants were presented with an Error and Confidence Judgments Rating Scale after their motor response. Here, the x-axis indicated a linear scale from visual onset to "+", representing the temporal position of the motor response relative to the visual stimulus onset, while the y-axis ranged from 0 to 100 % to denote confidence in this decision. After the participants provided their judgment, the next trial began within a random ITI of 1 to 3 s after marking their responses on the scale, potentially accompanied by feedback.

Participants began testing blocks with either the TP task or the RT task, with the order counterbalanced across participants to control for order effects. The TP task included two blocks, while the RT task included one block. This design was intended to ensure a larger dataset for the TP task.

3. Data preprocessing

Our analysis relied on custom-written R scripts (version 2023.12.1 + 402). During the data preprocessing phase, TPs below 0.5 s were excluded to ensure the inclusion of reliable responses for subsequent analysis.

Following these exclusions, z-score criteria were applied per subject to the TPs, RTs, and decision times (DTs) within each subject and block for both the TP and RT datasets. This standardization involved subtracting the mean (*M*) and dividing by the standard deviation (*SD*) of each variable within its respective condition. Subsequently, outliers beyond 3 standard deviations from the mean were filtered out. Originally, we had 5,078 trials for the time production (TP) task and 2,460 trials for the reaction time (RT) task. After excluding performance measures with z-scores exceeding 3 standard deviations, the dataset was reduced to 4,950 trials for the TP task and 2,360 trials for the RT task.

In the Error and Confidence Judgments Rating Scale, data scaling was essential for consistent and meaningful analysis. The scale was set within a 1200x1200 pixel square on a 2400x1350 pixel computer screen. The adjustments made ensure that the data is centered around zero and scaled to a consistent range, which can be particularly useful for visual comparisons and aligning different types of measurements within the same graphical representation. To characterize confidence levels within trials, Y-axis ratings were categorized into high and low-confidence groups based on a median split, facilitating straightforward comparative analysis across all trials. Additionally, a percentage confidence metric was derived to express Y-axis confidence values on a standardized 0 to 100 % scale. This transformation mapped confidence ratings onto a percentage scale by subtracting each Y-axis value from 1275 (a reference point)

and then scaling this difference to a percentage of 1200.

Overall, these preprocessing steps ensure uniform scaling of data across subjects and blocks while mitigating the potential bias of extreme values in subsequent analyses.

3.1. Statistical analysis

We employed Linear Mixed-Effect Models (LMMs) estimated via Restricted Maximum Likelihood (REML) to handle multiple observations per subject across experimental blocks. LMMs are well-suited for analyzing continuous response variables, such as TPs, RTs, Error and Confidence Judgment Scale ratings, and response delays observed in our study. These models account for both fixed effects, which represent consistent factors across experiments, and random effects, which capture variability within the population (Barr et al., 2013).

Following Kononowicz et al. (2022), our LMMs incorporated a nested random effect structure to account for performance variability within each subject. For analyses involving single-block and single-trial assessments, we began with models including only an intercept term and systematically evaluated random effects structures using the Chi-square test and Akaike Information Criterion (AIC). Ultimately, "subjects" were included as a random effect in the final models. Model parameters were standardized by fitting the models to a standardized dataset, and statistical significance was assessed using a Wald t-distribution approximation.

4. Results

4.1. Consistent Temporal Precision and Accuracy Across All Conditions in TP and RT tasks

Participants performed TP or RT tasks. At the beginning of each trial, they were instructed with a cue about the type of task (TP or RT) and handedness (DH or NDH). In the TP task, participants reproduced a target interval by either pressing a button twice (Press trials) or moving a joystick down and back to the center position twice (Move trials). In the RT task, participants reacted to a target stimulus by either pressing a button (Press trials) or moving a joystick downward to reach the white square (Move trials) as quickly as possible when it appeared at the bottom of the screen (Fig. 1B, right panel).

We first analyzed TP accuracy and precision in both Move and Press conditions. In the TP task, both Move and Press trials demonstrated high accuracy, with mean TPs closely matching the target duration. There were no significant differences between conditions, suggesting that participants have adapted to different motor demands in terms of time production performance (Fig. 2A, left panel; ß = −0.00099, 95 % CI [-0.05, 0.06], t (120) = -0.04, P = 0.972). Additionally, precision was high for both conditions, as indicated by small σ(TPs) values. Similar to TP accuracy, there were no significant differences between conditions in terms of their TP precision (Fig. 2A, right panel; ß = 0.00671, 95 % CI [-0.0074, 0.02], t (120) = 0.94, P = 0.350). In both conditions, single-block TP precision increased proportionally with μ(TPs), highlighting consistent TP properties regardless of variation in motor demand (Fig. 2C, bottom inset, left panel; ß = 0.12, 95 % CI [-0.49, 0.72], t (118) = 0.38, P = 0.708).

Moreover, both DH and NDH conditions demonstrated accurate time production, with μ(TPs) closely matching the target duration (Fig. 2C, left panel; ß = 0.00183, 95 % CI [-0.04, 0.04], t (120) = 0.08, P = *0*.933). Additionally, both conditions exhibited precise TPs (Fig. 2C, right panel; ß = −0.00317, 95 % CI [-0.02, 0.01], t (120) = -0.44, P = *0*.662), with σ(TPs) increasing proportionally with μ(TPs), highlighting consistent response properties regardless of handedness conditions (Fig. 2C, bottom inset, right panel; ß = 0.01, 95 % CI [-0.46, 0.48], t (118) = 0.05, P = *0*.962).

The analysis of RT accuracy and precision in both Move and Press conditions reveals several key findings. Both conditions exhibited high accuracy in RTs, with μ(RTs) closely aligning with the onset of the visual stimulus. However, Press RTs were shorter than Move RTs, likely due to the quicker response required to the visual stimulus (Fig. 2B, left panel; ß = −0.18, 95 % CI [-0.21, −0.15], t(56) = -12.54, P < 0.001). Button presses involve less movement and motor effort, leading to faster reactions, whereas joystick movements involve a more complex motor routine, resulting in longer RTs. Participants demonstrated precise performance, as reflected by small σ(RTs) values, with no significant difference observed between Move and Press conditions (Fig. 2B, right panel; ß = 0.02, 95 % CI [-0.02, 0.05], t (56) = 1.03, P = *0*.308). Notably, σ(RTs) increased proportionally with μ(RTs) in both conditions. However, a significant difference in slopes, particularly for Press trials, suggested an interaction effect (Fig. 2D, bottom inset, left panel; ß = −0.52, 95 % CI [-0.99, −0.05], t (56) = -2.20, P = *0*.032).

Further analysis extended to RT accuracy and precision in DH and NDH conditions. Both conditions demonstrated high accuracy in RTs, with μ(RTs) closely aligning with the onset of the visual stimulus, and no significant difference between conditions was found (Fig. 2D, left panel; ß = 0.02, 95 % CI [-0.0073, 0.04], t (56) = 1.36, P = *0*.178). Additionally, both DH and NDH conditions exhibited high precision (Fig. 2D, right panel; ß = 0.03, 95 % CI [-0.0049, 0.06], t (56) = 1.68, P = *0*.098), with σ(RTs) increasing proportionally with μ(RTs), highlighting consistent response properties regardless of motor signal associated with handedness (Fig. 2D, bottom inset, right panel; ß = −0.13, 95 % CI [-0.57, 0.32], t (56) = -0.57, P = *0*.573).

4.2. Properties of error Judgement in TP and RT tasks

We examined the influence of motor demand on EJ ratings in both the TP and RT tasks. EJ ratings in the Press condition were closer to the "target onset, 0(px)" compared to the Move condition (Fig. 3A, left panel: μ(Error Judgement(s)): β = -2.21, 95 % CI [-4.16, −0.25], t (120) = -2.24, P = 0.027; right panel: σ(Error Judgement(s)): β = -1.37, 95 % CI [-2.57, −0.17], t (120) = -2.27, P = 0.025). In the RT task, a similar pattern was observed with EJ ratings in the Press condition being closer to the "target onset, 0(px)" (Fig. 3B,

left panel: μ(Error Judgement(s)): β = -2.69, 95 % CI [-3.80, −1.58], t (56) = -4.85, $P < 0.001$; right panel: σ(Error Judgement(s)): β = -1.23, 95 % CI [-2.43, −0.03], t (56) = -2.05, $P = 0.046$). Additionally, the DS condition showed closer ratings to the "target onset, 0″ compared to the NDS condition in the TP task (Fig. 3C, left panel: μ(Error Judgement(s)): β = -2.91, 95 % CI [-4.46, −1.37], t (120) = -3.74, $P < 0.001$; right panel: σ(Error Judgement(s)): β = 0.45, 95 % CI [-0.70, 1.60], t (120) = 0.78, $P = 0.438$), and marginally in the RT task (Fig. 3D, left panel: μ(Error Judgement(s)): β = -1.05, 95 % CI [-2.12, 0.02], t (56) = -1.96, $P = 0.055$; right panel: σ(Error Judgement(s)): β = 0.12, 95 % CI [-1.10, 1.35], t (56) = 0.20, $P = 0.840$).

It is worth noting that while the Press and Move conditions did not significantly affect the precision of time production, this does not hinder the investigation of Error Judgement accuracy. The same statistical properties of TP provide an ideal context for examining how motor conditions influence error monitoring. Although the effects of induced motor signals on TP and RT were not directly observed, our findings suggest that motor signals modulate EJ ratings rather than directly contributing to TP variability.

4.3. Impact of motor signal on temporal error monitoring

To investigate self-monitoring of temporal errors, participants were asked to indicate their temporal errors and confidence using Error and Confidence Judgment Rating Scales (Fig. 4A). This scale utilized a coordinate system, with the x-axis representing duration or reaction time, and the y-axis representing confidence levels. Participants used a joystick during test trials to mark their judgments, with feedback provided in half of the cases. Two predictions were formulated based on the Clock and Motor hypotheses to explore the impact of motor signal on temporal error monitoring performance. The Clock Hypothesis suggests that inducing an additional motor signal would impair monitoring performance (Fig. 4B, left panel), while the Motor Hypothesis proposes that monitoring of temporal events is enhanced by an additional motor signal (Fig. 4B, right panel).

In the TP task, LMM analysis of the combined data from all conditions revealed a significant relationship between EM responses and TPs (β = 0.0054, 95 % CI [0.0049, 0.0059], t(4946) = 21.90, $p < 0.001$). Shorter TP trials elicited shorter EM responses in both Move (Fig. 4C, left panel, light red; β = 0.0049, 95 % CI [0.0042, 0.0055], t(4944) = 14.93, $P < 0.001$) and Press trials (Fig. 4C, left panel, dark red; β = 0.0061, 95 % CI [0.0054, 0.0068], t(4944) = 16.93, $P < 0.001$). In line with prediction 1 and the Clock Hypothesis, a significant slope difference between Press and Move conditions indicated more effective Error Judgements error monitoring in Press trials (Fig. 4C, left panel; β = 0.0012, 95 % CI [0.0003, 0.0021], t(4944) = 2.59, $P = 0.010$). These results highlight the substantial impact of motor signal on EM performance, despite its lack of influence on TP performance. The model comparison showed that the interaction model, which includes 'TP * motor demand', better explains variability in EM responses than the original model (ΔAIC = 4.6, $P = 0.013$). This indicates that accounting for the interaction between motor demand conditions improves the model fit and reveals differences in slopes.

In the DH condition (Fig. 4C, right panel, solid line; β = 0.0051, 95 % CI [0.0044, 0.0058], $t(4944) = 14.72$, $P < 0.001$) and NDH condition (Fig. 4C, right panel, dashed line; β = 0.0057, 95 % CI [0.0050, 0.0064], $t(4944) = 17.05$, $P < 0.001$), a significant correlation between actual TPs and EM responses on single trials was found, without any interaction effect (Fig. 4C, right panel; β = -0.0006, 95 % CI [-0.0015, 0.0003], $t(4944) = -1.27$, $P = 0.205$). This suggests that motor signal associated with handedness did not impact Error Judgements. The model comparison further supported this, indicating that the inclusion of the handedness factor in the baseline model did not improve the baseline model (ΔAIC = 0.2, $P = 0.122$).

In the RT Task, LMMs revealed a significant relationship between EM responses and actual RTs (β = 0.02, 95 % CI [0.01, 0.02], $t(2356) = 26.64$, $P < 0.001$). Significant relationships between EM responses and RTs were observed across all conditions: Move (Fig. 4D, left panel, light red; β = 0.01, 95 % CI [0.01, 0.01], $t(2354) = 16.98$, $P < 0.001$), Press (Fig. 4D, left panel, dark red; β = 0.02, 95 % CI [0.02, 0.02], $t(2354) = 20.89$, $P < 0.001$), DH (Fig. 4D, right panel, solid line; β = 0.02, 95 % CI [0.02, 0.02], $t(2354) = 22.60$, $P < 0.001$), and NDH (Fig. 4D, right panel, dashed line; β = 0.01, 95 % CI [0.01, 0.02], $t(2354) = 17.76$, $P < 0.001$). Significant interaction effects were observed for the Move and Press conditions, indicating the impact of motor demands on RTs, given the task's

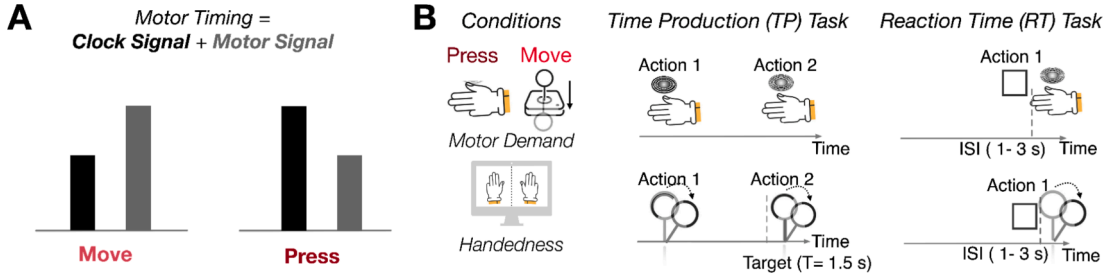


Fig. 1. Hypotheses and predictions based on the two-component clock model, consisting of clock and motor components.
A. Two panels depict the relative contribution of clock and motor components in Press and Move conditions.
B. Press and Move conditions in the TP and RT tasks. During the TP task, participants were instructed to accurately reproduce a target interval by pressing a button twice (Press trials) or moving a joystick down and back to the center position twice (Move trials). In the RT task, they were required to react to a target stimulus by pressing a button (Press trials) or moving a joystick inside the stimulus (Move trials) as quickly as possible upon its appearance at the bottom of the screen. Each trial included instructions for performing the task in a Press or Move condition and specified whether to use their DH or NDH based on cue information presented at the trial's onset.

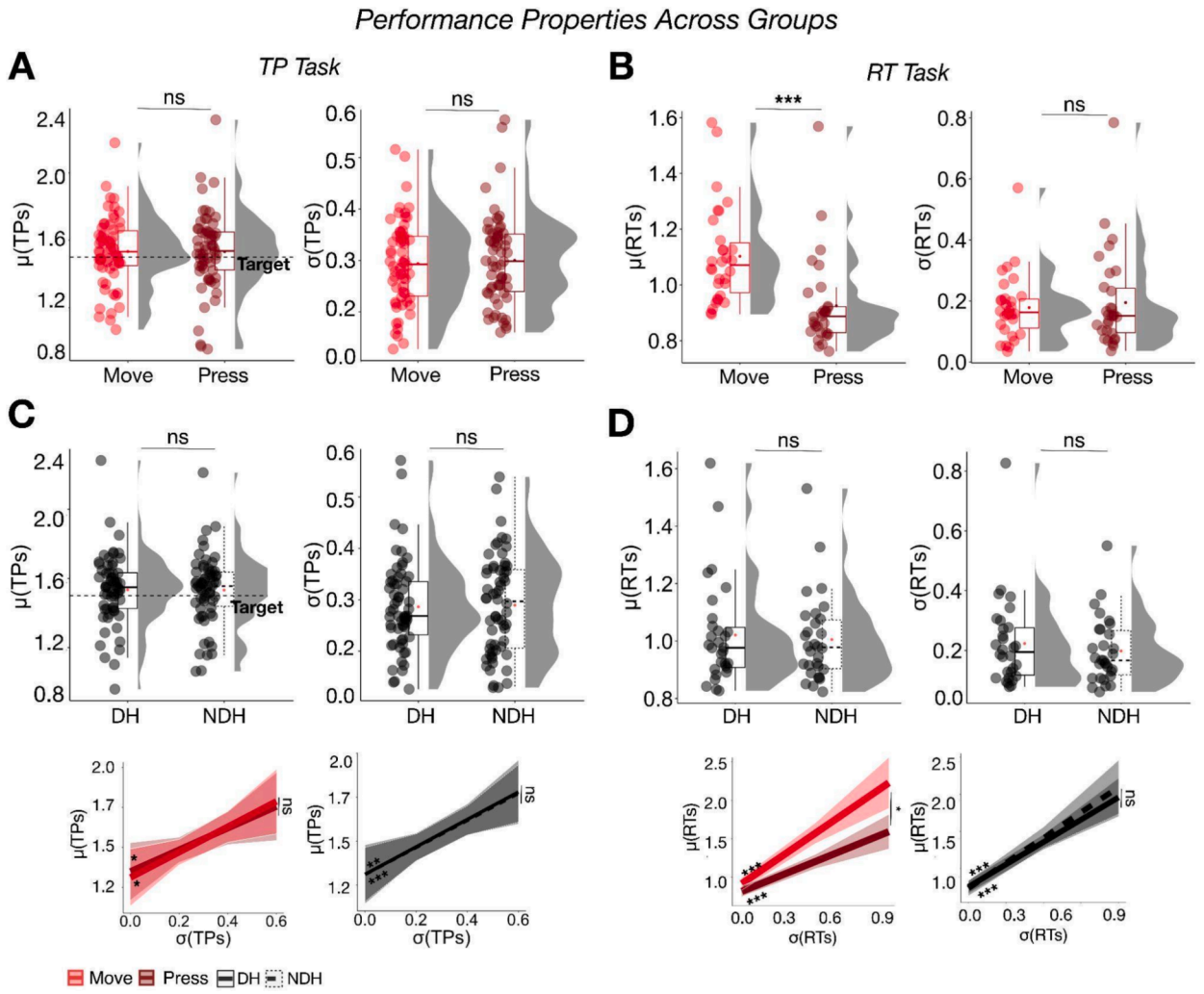


Fig. 2. Humans demonstrated accurate and precise TP and shorter and precise RT intervals on a single-trial basis, regardless of motor conditions, highlighting the consistent preservation of TP and RT properties.

A-C. TP accuracy and precision across different conditions. Move trials (depicted in light red) and Press trials (shown in dark red) were analyzed using box plots with density curves to illustrate the mean (M = μ) (A, left panel) and precision (SD = σ, standard deviation) (A, right panel) of TPs for both conditions at a single block-level. The significance of the differences between conditions is indicated by asterisks above the plots: ***, **, *, "marginal", and "ns" denote varying levels of statistical significance as < 0.001, <.01 < 0.05, = 0.05, > 0.05 respectively. The dots within each boxplot represent the mean values for each subject's blocks within their respective conditions. Participants consistently demonstrated accurate TPs in both conditions, with mean TPs closely aligning with the target duration. Similarly, in both DH (C, depicted in solid line) and NDH (C, shown in dashed line) conditions, participants produced accurate TPs. The lower panels depict the relationship between mean and precision in the TP task.

B-D. RT mean and precision across different conditions. The plots depict the mean and precision of RTs across Move-Press (panel B), and DH-NDH (panel D) conditions. The bottom panels illustrate the association of mean and precision in the RT task.

motor performance requirement (Fig. 4D, left panel; β = 0.0057, 95 % CI [0.0037, 0.0077], $t(2354) = 5.51$, $P < 0.001$). Additionally, an interaction effect was found for the handedness conditions, resulting in decreased EM performance (a weaker relationship between EJ ratings and RTs) as indicated by a less steep slope (Fig. 4D, bottom right panel; β = 0.0045, 95 % CI [0.0025, 0.0066], t(2354) = 4.28, p < 0.001). These findings were further supported by model comparisons between baseline and more complex models including motor demands (ΔAIC = 282.65, $P < 0.001$) and handedness (ΔAIC = 27.6, $P < 0.001$).

To assess how well participants align their Error Judgements with actual performance, we calculated the absolute distance between the z-scored values of the EJ ratings and actual performance. This method accounts for the different scales of these measures at the single-trial level. We hypothesized that a positive slope between trial performance and EJ ratings indicates better monitoring. Therefore, the distance between these measures could serve as an index of trial-level alignment, where smaller distances reflect better alignment and, consequently, more accurate EM. We expected that groups with lower motor signal values would show smaller distances.

Our results reveal that the Press group exhibited smaller coefficient values, suggesting better EJ alignment with single trial

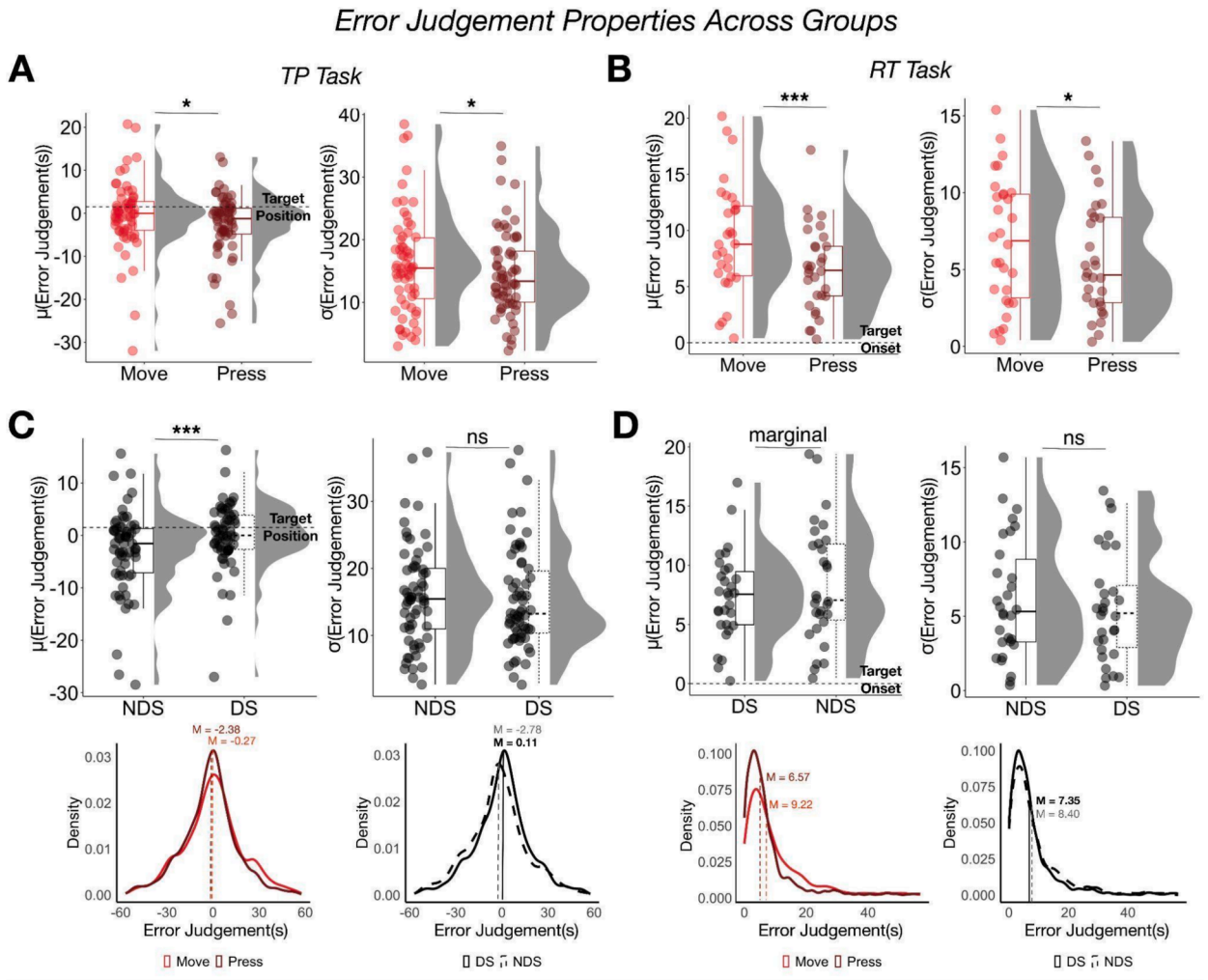


Fig. 3. Properties of Error Judgement (EJ) across conditions.

A-C. EJ properties in TP task. The panels represent the mean and standard deviation of EJ ratings across conditions in the TP task. Move trials (depicted in light red) and Press trials (shown in dark red) are depicted using box plots with density curves to illustrate the mean (M = μ) (left panels) and precision (SD = σ, standard deviation) (right panels) of EJ ratings for both conditions at a single block level.

Mean EJ ratings in the Press condition were more concentrated around the "target position, 0 (px) (pixels, middle of the rating scale)," compared to the Move condition (panel A). The DS condition also exhibited more concentrated EJ ratings around the "target position, 0 (px)," compared to the NDS condition (panel C). EJ distribution was narrower for the Press than the Move condition, suggesting that our manipulation is reflected in EJ rather than in TP. No significant distribution width differences were observed concerning handedness.

B-D. EJ properties in RT task. The panels represent the mean and precision of EJ ratings in the RT task across Move-Press (panel B) and DH-NDH (panel D) conditions. A similar pattern is observed, with EJ ratings in the Press condition being closer to the "target onset, 0 (px)," compared to the Move condition. The DS condition also displayed marginally closer EJ ratings to the "target onset, 0 (px)" than the NDS condition. In parallel with the TP task, the Press group's distribution was narrower compared to the Move group, but this effect was absent in handedness conditions. The density in panels C and D further illustrates the distribution of EJ ratings across the TP and RT tasks.

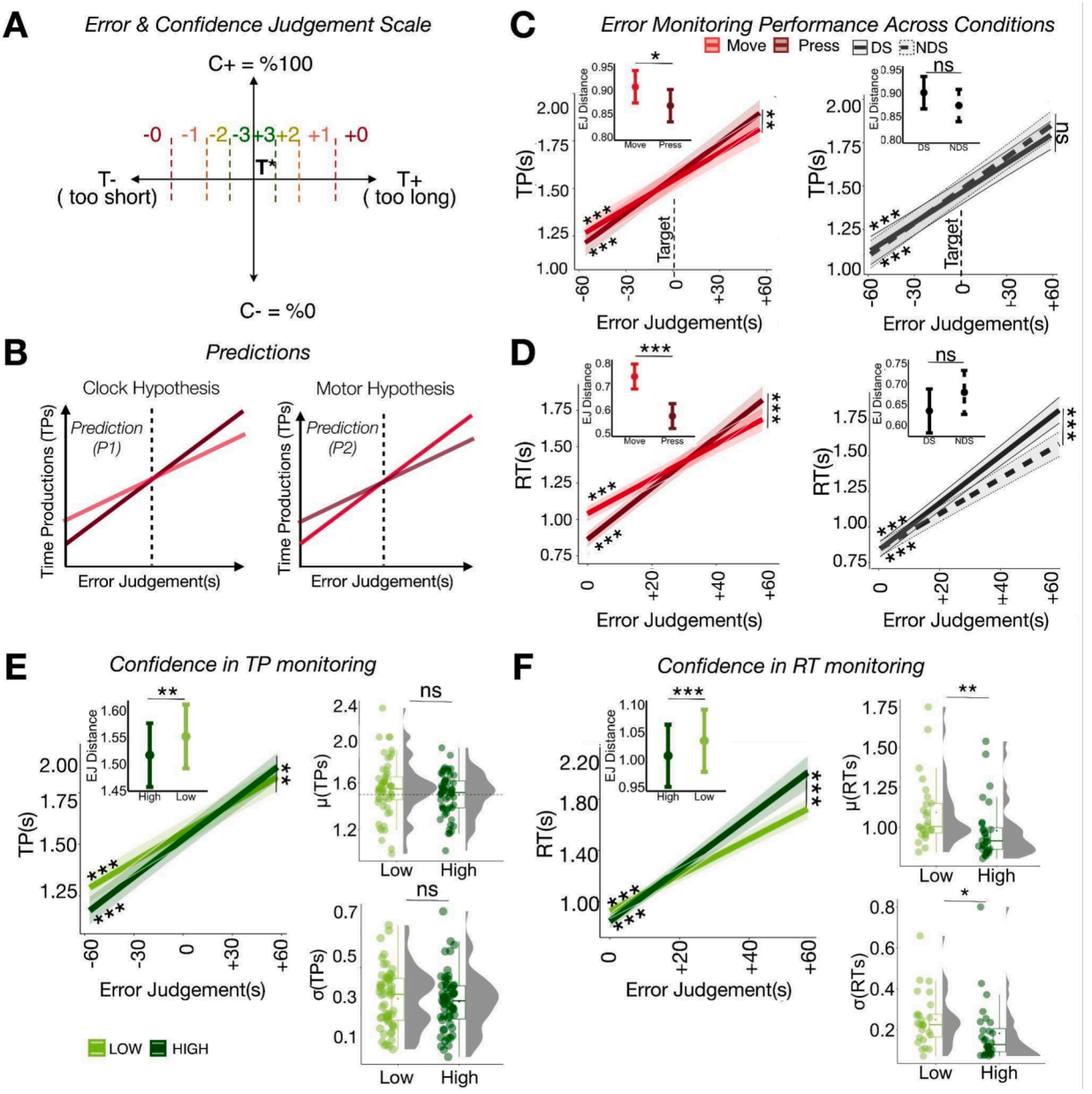
performance, as the distance between the EJ ratings and actual performance was smaller compared to the Move group in both the TP (Fig. 4C, right inset panel, β = -0.04, 95 % CI [-0.08, −1.65e-03], t (4946) = -2.04, $P = 0.040$) and RT tasks (Fig. 4D, right inset panel, β = -0.17, 95 % CI [-0.22, −0.12], t (2356) = -6.51, $P < 0.001$). However, this pattern was absent among the Handedness conditions, both in the TP task (Fig. 4C, left inset panel, β = -0.03, 95 % CI [-0.07, 0.01], t (4946) = -1.39, $P = 0.165$) and the RT task (Fig. 4D, left inset panel, β = 0.05, 95 % CI [-6.14e-03, 0.10], t (2356) = 1.73, $P = 0.084$).

4.4. Impact of single-trial confidence on temporal error monitoring and performance metrics

In addition to the previous analysis, LMMs were used to examine the relationship between participants' confidence decisions regarding their EM responses and their performance in both TP and RT Tasks. This analysis categorized decision confidence conditions

based on a median split of decision confidence ratings on the y-axis to distinguish between low and high confidence levels. The results indicated that on single trials where participants rated their confidence in their EM responses after test trials, higher confidence levels were associated with better EM performance (a stronger relationship between EJ ratings and TPs) compared to low confidence trials for both TP (Fig. 4E, main panel; $\beta = 0.0014$, 95 % CI [0.0005, 0.0023], $t(4944) = 2.94$, $P = 0.003$) and RT Tasks (Fig. 4F, main panel; $\beta = 0.0064$, 95 % CI [0.0037, 0.0090], $t(2354) = 4.72$, $P < 0.001$). This suggests that participants were able to monitor their errors and make informed confidence decisions based on their performance. Furthermore, these results were supported by model comparison analysis (TP task confidence: $\Delta AIC = 4.6$, $P = 0.013$; RT Task confidence: $\Delta AIC = 24.44$, $P < 0.001$). Notably, in both TP (Fig. 4E, main panel, high-confidence: $\beta = 0.0061$, 95 % CI [0.0054, 0.0068], $t(4944) = 17.89$, $P < 0.001$; low-confidence: $\beta = 0.0047$, 95 % CI [0.0040, 0.0054], $t(4944) = 13.87$, $P < 0.001$) and RT Tasks (Fig. 4F, main panel, high-confidence: $\beta = 0.02$, 95 % CI [0.02, 0.02], $t(2354) = 16.91$, $P < 0.001$; low-confidence: $\beta = 0.01$, 95 % CI [0.01, 0.02], $t(2354) = 21.01$, $P < 0.001$), participants demonstrated effective Error Judgements regardless of confidence levels.

Parallel to the previous findings on EJ Distance, our results reveal that the High Confidence group exhibited smaller coefficient



(caption on next page)

Fig. 4. Induced motor signal and single-trial confidence effect on Temporal Error Judgement during TP and RT tasks.
A. Error and Confidence Judgment Rating Scale. To assess participants' ability to report temporal errors in their TPs and express confidence in their monitoring responses, the Error and Confidence Judgment Rating Scale was given after each trial. This scale employed a coordinate system where the x-axis represented a linear scale from the shortest to the longest duration for the TP task and from the shortest to the longest reaction times for the RT task. Simultaneously, the y-axis ranged from 0 % to 100 % which indicates explicit confidence in monitoring decisions. Participants used a joystick to mark the intersection point of their judgments on both temporal error and confidence for both TP and RT tasks. Feedback on monitoring performances was provided in 50 % of the trials.
B. Investigating the impact of motor signal on temporal error monitoring: Predictions based on the Clock and Motor hypotheses. Two predictions were formulated based on the Clock Hypothesis and the Motor hypothesis. The Clock Hypothesis (left panel) indicates that an additional signal would impair temporal error monitoring performance if TP error relies on an introspective assessment of the clock signal. Consequently, the slope for the Press trials (dark red line) would be higher than that for the Move trials (red line). Alternatively, the Motor hypothesis (right panel) posits that if temporal error monitoring is based on the monitoring of motor signal, the slope in the Move trials will be higher than that of the Press trials, as error monitoring of temporal events is influenced by induced motor signal. According to the framework of the Clock and Motor hypotheses, these predictions were formulated to further test how motor signal impacts temporal error monitoring performance.
C. Motor demands and handedness effects on temporal error monitoring in TP tasks. All panels depict the predicted fits derived from LMM models (see Supp. Mat. 2 for actual data points). Participants' EM responses were determined by the position on the EM rating scale where they marked their responses. The center of the scale corresponded to the target time of 1.5 s, while the extreme left and right positions represented the shortest and longest times, respectively. Confidence intervals, depicted as shaded areas, were plotted for each condition at a population level. The width of the shadings represents population-level estimates by considering uncertainty in the signal parameters. In shorter TP trials, participants' monitoring responses indicated lower pixel values, reflecting their ability to detect temporal errors on a single trial basis in both Move and Press conditions (left panel). A higher slope of the linear function suggested an increased EM response with current-trial TP. Notably, this slope difference between the Press and Move conditions was significant, indicating that participants in Press trials monitored their errors much better. This was evident by the stronger correlation of both variables relative to the Move trials (left panel).
Similarly, a robust relationship between ongoing trial TPs and participants' EM responses was observed in the handedness conditions (right panel). This suggests that participants were able to detect temporal errors on a single-trial basis. Overall, the findings highlight the influence of movement demands on temporal error monitoring performances, and participants' ability to accurately assess temporal errors.
The upper inset panels represent the trial-by-trial distance between z-scored Error Judgments (EJ) and current TP estimates, defined as EJ Distance. We hypothesized that groups with lower motor signal values would exhibit smaller EJ distances, indicating better alignment between EJ and actual performance. The results show that the Press group demonstrates a smaller distance compared to the Move group, suggesting that participants in the Press group actively monitored and adjusted their EJ ratings to minimize the discrepancy between perceived Error Judgement and true performance. However, this pattern was not observed among the Handedness groups, which is consistent with the results of the EM and actual performance analysis.
D. Motor demands and handedness effects on temporal error monitoring in RT task. Both figures depict the predicted fits derived from LMM models (see Supp. Mat. 2 for actual data points). Both Press (left panel) and DH (right panel) trials exhibited a better slope than the alternative conditions with current RTs on a single trial. This indicates that the monitoring of RTs was influenced by the induced motor signal between conditions. The results suggest that since RT tasks rely on monitoring RT onset relative to the visual stimulus and predominantly depend on motor performance, they may be more susceptible to manipulations. Overall, in both RT and TP tasks, the findings support predictions of the Clock Hypothesis, indicating that inducing motor signal would impair temporal error monitoring.
The upper inset panels for the RT task show results similar to the TP task (panel C). The Press group demonstrates a smaller distance compared to the Move group. However, this pattern was not observed among Handedness groups.
E, F. (main panels) confidence impacts on error monitoring in TPs and RTs. The analysis investigated correlations at different confidence levels by categorizing trials as "High" and "Low" based on a median split of confidence ratings on EM responses (y-axis of the response scale). All panels present predicted fits derived from LMM models (see Supp. Mat. 2 for actual data points). Results showed that Error Judgement ratings were higher in high-confidence trials compared to low-confidence trials for both TP and RT measures. The relationship between EM responses, TPs, and RTs at different confidence levels was also analyzed to assess correlations between participant confidence and EM responses across all motor signal conditions separately (see Supp. Mat. 1 for details). Higher confidence in monitoring decisions was associated with improved monitoring accuracy, demonstrated by steeper slopes.
The upper inset panels for the TP and RT tasks show results similar to the previous analysis (panel C&D). The high-confidence group demonstrates a smaller distance compared to the low-confidence group.
E, F. (inset panels) Confidence impacts performance metrics of TPs and RTs. For TP tasks, confidence levels did not significantly affect performance metrics, as there were no significant differences between high and low-confidence trials. However, in the RT task, high-confidence trials demonstrated both higher precision (lower σ(RTs)) (F, bottom inset panels) and accuracy (lower μ(RTs)) (F, upper inset panels) compared to low-confidence trials. These findings suggest a close association between participants' confidence levels and their actual performance in the RT task, suggesting the influence of self-motor control on confidence judgments and monitoring. Conversely, in TP tasks, confidence levels did not seem to correlate with performance metrics (E, inset panels). This highlights that participants might have prioritized Error Judgement over considering their performance during timing tasks, or they might not have had access to performance metrics even though they were able to monitor their errors, suggesting an intrinsic temporal error monitoring mechanism.

values, suggesting better EJ alignment with actual performance, as the distance between the EJ ratings and actual performance was smaller compared to the Low Confidence group in both the TP (Fig. 4E, upper inset panel, $\beta = 0.03$, 95 % CI [0.01, 0.06], $t(4944) = 2.85$, $P = 0.004$) and RT tasks (Fig. 4F, upper inset panel, $\beta = 0.08$, 95 % CI [0.05, 0.11], $t(2354) = 4.80$, $P < 0.001$).

Finally, we investigated the impact of confidence on performance metrics in both TP and RT Tasks. In the TP task, confidence levels did not significantly influence performance metrics, as there were no substantial differences found between high and low-confidence trials (Fig. 4E, upper inset panel (accuracy): $\beta = -0.05$, 95 % CI [-0.11, 0.004], $t(107) = -1.84$, $P = 0.068$; bottom inset panel (precision): $\beta = -0.01$, 95 % CI [-0.03, 0.008], $t(107) = -1.23$, $P = 0.221$). Conversely, in the RT Task, high-confidence trials exhibited higher

precision and accuracy compared to low-confidence trials (Fig. 4F, upper inset panel (accuracy): β = -0.11, 95 % CI [-0.18, −0.04], *t* (50) = -3.03, *P* = *0*.004; bottom inset panel (precision): β = -0.06, 95 % CI [-0.11, −0.02], *t*(50) = -2.73, *P* = *0*.009). This suggests a strong relationship between participants' confidence levels and their actual performance in the RT Task, indicating that self-motor control significantly influences confidence judgments and monitoring.

We examined how decision times (DT) and confidence ratings were affected by motor demands. The figure presents predicted values derived from LMMs for confidence and decision times across different conditions in TP and RT Tasks. Each subplot displays mean values with error bars indicating the standard error of the mean. Conditions are grouped into columns: Move trials (light red) vs. Press trials (dark red), and DH trials (solid line) vs. NDH trials (dashed line). DTs were measured as the time interval between the onset of the scale presentation on the screen and the participant's response for each trial. Confidence ratings were normalized to percentages to standardize the measurements.

In both TP and RT tasks, the press and dominant hand trials consistently exhibited higher confidence levels and shorter decision times than the move and nondominant hand trials. This pattern suggests that the type of motor action and handedness significantly influence both confidence levels and decision times. As the induced signal decreased, confidence in monitoring responses increased, and decision times to respond on the scale were shortened.

4.5. *Induced motor signal impacts confidence levels and decision times*

In exploring participants' Error Judgement performance between Press and Move trials, we aimed to ascertain whether this difference extended to other facets of the decision-making process. Enhanced Error Judgement typically correlates with a higher rate of information accrual, indicating that individuals gather evidence more rapidly per unit of time (Desender et al., 2021; Pereira et al., 2022). Therefore, we hypothesized that conditions demonstrating superior Error Judgement would also exhibit shorter decision times and higher confidence levels, reflecting quicker and more assured decision-making processes. LMMs were employed to analyze decision times (seconds), defined as the duration from scale onset to participant response, and confidence ratings across various experimental conditions.

Fig. 5A and 5B depict the impact of increased motor signal on confidence levels and decision times across TP and RT Tasks. LMM results revealed that in both tasks, confidence levels were significantly higher in Press trials compared to Move trials (Fig. 5A, upper left panel (TP task): β = 0.89, 95 % CI [0.24, 1.55], t(4946) = 2.68, *P* = 0.007; Fig. 5B, upper left panel (RT Task): β = 3.33, 95 % CI [2.44, 4.23], t(2356) = 7.29, *P* < 0.001). Similarly, confidence levels were significantly higher in DH trials compared to NDH trials in both tasks (Fig. 5A, upper right panel (TP task): β = 0.77, 95 % CI [0.12, 1.43], t(4946) = 2.32, P = 0.020; Fig. 5B, upper right panel (RT task): β = 2.34, 95 % CI [1.44, 3.25], t(2356) = 5.10, *P* < 0.001).

Consistent with our prediction, decision times were shorter in Press trials compared to Move trials in both tasks (Fig. 5A, bottom left panel (TP task): β = -0.21, 95 % CI [-0.30, −0.11], t(4946) = -4.31, *P* < 0.001; Fig. 5B, bottom left panel (RT task): β = -0.16, 95 % CI [-0.29, −0.03], t(2356) = -2.48, *P* = 0.013). Additionally, while there were no significant differences in decision times between DH and NDH trials in the TP task (Fig. 5A, bottom right panel: β = -0.07, 95 % CI [-0.16, 0.03], t (4946) = -1.41, P = 0.157), a significant difference was observed in the RT task (Fig. 5B, bottom right panel: β = -0.21, 95 % CI [-0.34, −0.08], t(2356) = -3.20, *P* = 0.001).

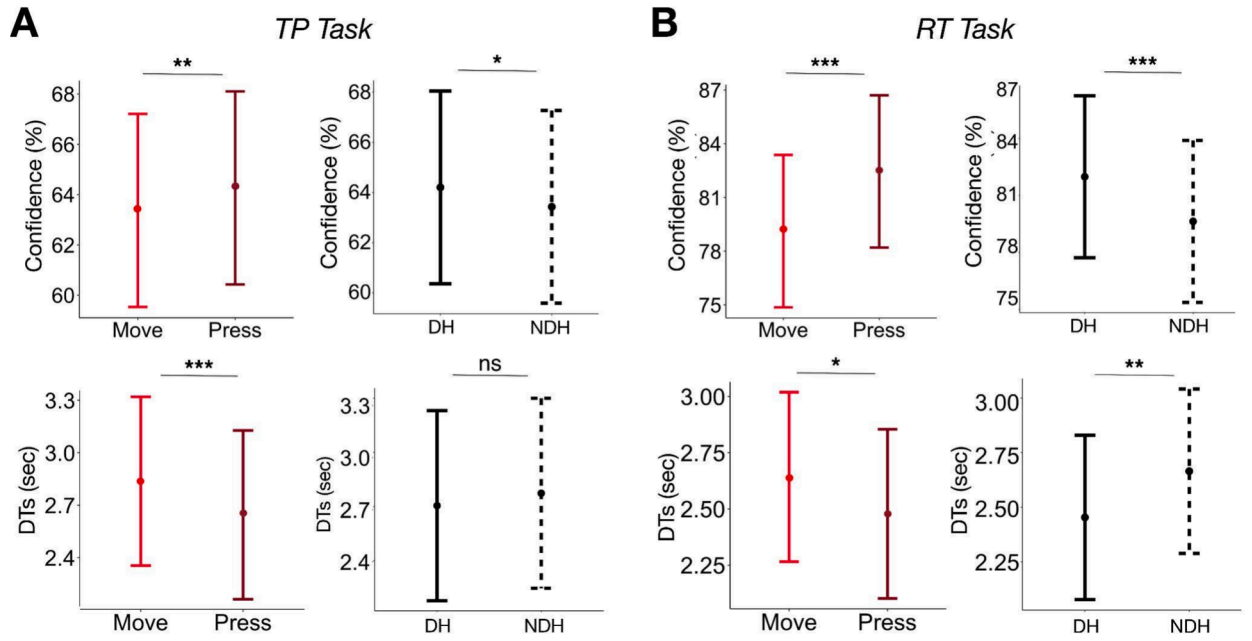


Fig. 5. A larger motor signal decreases decision confidence and increases decision times.

These findings align with previous analyses reported in Fig. 4. Reduced motor signal in the Press condition facilitated more efficient Error Judgement in both the TP and RT tasks, whereas reduced signal in the DH condition only influenced Error Judgement in the RT task. Consequently, participants exhibited faster decision times in the Press trials across both tasks and in the DH trials specifically in the RT task. Participants showed higher confidence in their decisions under both the Press and DH trials in both tasks. This suggests that while motor signal induced by handedness might not disrupt Error Judgement in the TP task, it may be reflected in confidence judgments. Overall, these results emphasize the significant influence of motor action type and handedness on confidence levels and decision times in tasks requiring precise temporal and motor responses.

5. Discussion

Temporal error monitoring has been defined as the ability to correctly report one's errors without relying on external feedback. Studies suggested that subjects have some degree of access to their noise in interval timing. Even more remarkably, they seem to be cognitively aware of their own temporal errors in self-generation of time intervals. Our goal was to assess what kind of information subjects consider when evaluating temporal errors during self-assessment. We considered two sources of signal in interval timing (Wing & Kristofferson, 1973): motor and clock signal. To that end we investigated how motor signal affects temporal error monitoring in time production (TP) and reaction time (RT) tasks. We formulated predictions based on two contrasting hypotheses: according to the Clock Hypothesis, error monitoring performance decreases with increased clock signal (noise), whereas the Motor Hypothesis suggests that motor variability might improve error monitoring by creating a reference point between the intended and actual motor actions, which helps identify errors on a trial-by-trial basis. Our findings supported the Clock Hypothesis, showing that increasing motor demands decreased the accuracy of temporal error monitoring performance in both TP and RT tasks. While handedness did not decrease monitoring performance and decision times, this may be because the handedness condition itself did not induce sufficient signal to affect these metrics. However, participants' confidence levels were still sensitive to variations induced by motor differences, regardless of their magnitude.

In addition to evaluating metacognitive accuracy in temporal error monitoring, our study investigated how induced motor signal influences retrospective decision confidence. Note that participants were asked to report confidence in their temporal error assessment rather than confidence in temporal production. We observed a significant increase in confidence ratings under the Press condition in both TP and RT tasks. In line with the modulation of retrospective confidence, there was a decrease in decision times (DT) when the motor signal was lower. Together, these results provide evidence for improved temporal error reporting under low motor signal, supporting the claim that participants monitor clock signal in temporal error monitoring.

In line with the Clock Hypothesis, Pavallier et al. (2021) investigated introspective evaluations of reaction times under varying conditions. Using EMG recordings, the study separated introspective reaction times into premotor time, related to decision-making, and motor time, associated with motor execution. The results demonstrated that participants could access and evaluate both decision and motor execution durations. Similar to the findings of Pavailler et al. (2024), although our results align with the Clock Hypothesis, they do not rule out the possibility that participants access motor signals during temporal error monitoring. The primary strength of our experimental setup lies in its ability to demonstrate the relative contributions of these factors and their importance in temporal error monitoring.

The impact of motor action on ongoing decisions has been studied in decision-making paradigms, where it was hypothesized that the formation of confidence judgments may interact with different types of motor activity (Fleming et al., 2015; Fleming & Daw, 2017; Gajdos et al., 2019). Our paradigm differs from previous studies investigating the impact of motor actions on decision-making (Sanchez et al., 2024). Here, we manipulated motor demands throughout the performance of the task to be monitored, whereas previous studies manipulated motor demands at the level of the confidence report. Moreover, in temporal error monitoring paradigms participants are asked to report signed and continuous errors in their own behavior which stand out in the context of two alternative forced choice tasks.

Nonetheless, these studies highlight the monitoring hypothesis (Sanchez et al., 2024), proposing that action monitoring during response selection can influence retrospective confidence. In line with that hypothesis, one could speculate that an increase in action monitoring would increase confidence due to the larger involvement of action control over one's motor responses. In the current task, increased motor demand could be equated to an increase in action monitoring and thus predict higher confidence reports in the move condition. However, we found higher confidence in participants' error estimation in the press condition, suggesting that the observed effects cannot be attributed to the modulation of cognitive factors other than motor signal.

It is plausible to argue that confidence in error monitoring might vary by condition. Participants may monitor their performance more closely when using their non-dominant hand (NDH) or in tasks perceived as more challenging, while exerting less effort in easier conditions, such as with their dominant hand (DH). However, our data showed the opposite pattern: confidence ratings were higher in the DH condition compared to the NDH condition, suggesting that the proposed scenario is unlikely.

To extend the previous metrics used in temporal error monitoring, we examined how participants' confidence ratings on the probability of correctness of error-monitoring responses correlated with their actual monitoring performance. Prior research has predominantly treated confidence as a secondary evaluation of initial timing accuracy (Akdogan & Balci, 2017; Fromer et al., 2021). This phenomenon, known as metacognitive accuracy, interprets confidence judgments as reflecting subjective assessments of the likelihood of accurate timing behavior (Akdogan & Balci, 2017; Frith, 2021). To evaluate awareness levels regarding the accuracy of temporal error judgments, we instructed participants to rate their confidence in their temporal error judgments. Thus, these confidence reports constitute a higher-level confidence estimate, as they integrate time production error with the self-reported error, possibly by integration of posterior probabilities of both judgments (Li & Ma, 2020). We found that higher confidence levels in Error Judgments

were associated with more accurate temporal error monitoring performance in both TP and RT tasks across nearly all conditions. Interestingly, we found the lowest accuracy of temporal error judgments in the move condition under a low level of confidence. Thus, this new metric also furnishes the evidence for the main hypothesis assessed in the study.

To summarize, the current study provides valuable insights into the relationship between motor demands, temporal error monitoring performance, and confidence judgments. As motor demands can alter temporal error monitoring, our findings emphasize that the clock signal is the primary signal monitored by participants. This further suggests that participants evaluate temporal representations in temporal error monitoring, supporting the notion of temporal metacognition (Kononowicz et al., 2019; van Wassenhove, 2009) and metacognitive readout of timing variable (Kononowicz & van Wassenhove, 2019). These are further in line with the metacognitive evaluation of temporal judgment as manifested in the modulation of higher-order confidence judgment.

Author contributions

SNB and TWK conceived and planned the experiments. SNB carried out the experiments. SNB and TWK planned and carried out the statistical analyses. SNB prepared the first version of the manuscript.

CRediT authorship contribution statement

Sena N. Bilgin: Writing – original draft, Methodology, Data curation, Conceptualization. **Tadeusz W. Kononowicz:** Writing – review & editing, Writting - original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Funding

This work was supported thanks to OPUS grant 2019/35/B/HS6/04389 to TWK from the National Center of Science in Poland.

TWK and SB developed the research concept and study design. Coding was performed by SB. Testing and data acquisition was performed by SB. The analyses were performed by SB and TWK. SB and TWK wrote the first draft. SB and TWK wrote and revised the current version of the manuscript. SB and TWK reviewed and approved the final version of the manuscript.

Acknowlegments

This project has received funding from National Center of Science in Poland (grant nr: 2019/35/B/HS6/04389 to TWK). We thank Valérie Doyère and Bruce Brown for numerous enriching discussions about the study and for their feedback on the previous version of the manuscript. We would like to thank the anonymous referees for their valuable comments that improved the quality of the paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.concog.2025.103849.

Data availability

Data will be made available on request. The analysis code is avaible at https://github.com/xenablgn/opus-2023-motor-experiment.git.

References

Akdoğan, B., & Balcı, F. (2017). Are you early or late?: Temporal error monitoring. *Journal of Experimental Psychology: General, 146*(3), 347–361. https://doi.org/10.1037/xge0000265

Bader, F., & Wiener, M. (2021). Awareness of errors and feedback in human time estimation. *Learning & Memory, 28*(5), 171–177. https://doi.org/10.1101/lm.053108.120

Bader, F., & Wiener, M. (2024). Neuroimaging Signatures of Metacognitive Improvement in Sensorimotor Timing. *The Journal of neuroscience: the official journal of the Society for Neuroscience, 44*(9), Article e1789222023. https://doi.org/10.1523/JNEUROSCI.1789-22.2023

Balci, F., Freestone, D., Simen, P., Desouza, L., Cohen, J. D., & Holmes, P. (2011). Optimal temporal risk assessment. *Frontiers in Integrative Neuroscience, 5*, 56.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision, 10*(4), 433–436.

Bratzke, D., & Bryce, D. (2016). Temporal discrimination of one's own reaction times in dual-task performance: Context effects and methodological constraints.". *Attention, Perception, & Psychophysics, 78*, 1806–1816.

Corallo, G., Sackur, J., Dehaene, S., & Sigman, M. (2008). Limits on introspection: Distorted subjective time during the dual-task bottleneck. *Psychological science, 19*(11), 1110–1117. https://doi.org/10.1111/j.1467-9280.2008.02211.x

Desender, K., Donner, T. H., & Verguts, T. (2021). Dynamic expressions of confidence within an evidence accumulation framework. *Cognition, 207*, 104–522.

De Kock, R., Gladhill, K. A., Ali, M. N., Joiner, W. M., & Wiener, M. (2021). How movements shape the perception of time. *Trends in Cognitive Sciences, 25*(11), 950–963. https://doi.org/10.1016/j.tics.2021.08.002

Duyan, Y. A., & Balcı, F. (2019). Metric error monitoring in the numerical estimates. *Consciousness and Cognition, 67*, 69–76.

Duyan, Y. A., & Balcı, F. (2020). Monitoring line length reproduction errors. *Consciousness and cognition, 77*, 102831.

Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-Specific Disruption of Perceptual Confidence. *Psychological Science, 26*(1), 89–98. https://doi.org/10.1177/0956797614557697

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychol Rev, 124*(1), 91–114. https://doi.org/10.1037/rev0000045

Frith, C. D. (2021). The neural basis of consciousness. *Psychological Medicine, 51*(4), 550–562. https://doi.org/10.1017/S0033291719002204

Frömer, R., Nassar, M. R., Bruckner, R., Stürmer, B., Sommer, W., & Yeung, N. (2021). *Response-based outcome predictions and confidence regulate feedback processing and learning., eLife, 10*, Article e62825. https://doi.org/10.7554/eLife.62825

Gajdos, T., Fleming, S. M., Garcia, M. S., Weindel, G., & Davranche, K. (2018). Revealing subthreshold motor contributions to perceptual confidence. *Neuroscience of Consciousness, 2019*. https://doi.org/10.1093/nc/niz001

Getty, D. J. (1976). Counting processes in human timing. *Perception & Psychophysics, 20*(3), 191–197.

Kononowicz, T. W., Roger, C., & Wassenhove, V. (2019). Temporal metacognition as the decoding of self-generated brain dynamics. *Cerebral cortex (New York, N.Y.: 1991), 29*(10), 4366–4380. https://doi.org/10.1093/cercor/bhy318

Kononowicz, T. W., & van Wassenhove, V. (2019). Evaluation of Self-generated behavior: Untangling metacognitive readout and error detection. *Journal of cognitive neuroscience, 31*(11), 1641–1657.

Kononowicz, T. W., van Wassenhove, V., & Doyère, V. (2022). Rodents monitor their error in self-generated duration on a single trial basis. *Proceedings of the National Academy of Sciences of the United States of America, 119*(9), Article e2108850119. https://doi.org/10.1073/pnas.2108850119

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature neuroscience, 13*(8), 1020–1026.

Jovanovic, L., López-Moliner, J., & Mamassian, P. (2021). Contrasting contributions of movement onset and duration to self-evaluation of sensorimotor timing performance. *European Journal of Neuroscience, 54*(3), 5092–5111.

Li, H. H., & Ma, W. J. (2020). Confidence reports in decision-making with multiple alternatives violate the Bayesian confidence hypothesis. *Nature communications, 11*(1), 2004.

Macar, F., Vidal, F., & Casini, L. (1999). The supplementary motor area in motor and sensory timing: Evidence from slow brain potential changes. *Experimental brain research, 125*, 271–280.

Mamassian, P. (2008). Overconfidence in an objective anticipatory motor task. *Psychological Science, 19*(6), 601–606.

Marti, S., Sackur, J., Sigman, M., & Dehaene, S. (2010). Mapping introspection's blind spot: Reconstruction of dual-task phenomenology using quantified introspection. *Cognition, 115*(2), 303–313. https://doi.org/10.1016/j.cognition.2010.01.003

Öztel, T., Eskenazi, T., & Balcı, F. (2021). Temporal error monitoring with directional error magnitude judgments: A robust phenomenon with no effect of being watched. *Psychological Research, 85*, 2069–2078. https://doi.org/10.1007/s00426-020- 01379-0

Öztel, T., & Balcı, F. (2023). Humans can monitor trial-based but not global timing errors: Evidence for relative judgements in temporal error monitoring. *Quarterly Journal of Experimental Psychology, 76*(9), 2155–2163. https://doi.org/10.1177/17470218221145314

Öztel, T., & Balcı, F. (2024). Metric error monitoring as a component of metacognitive processing. *European Journal of Neuroscience, 59*(5), 807–821. https://doi.org/10.1111/ejn.16182

Paton, J. J., & Buonomano, D. V. (2018). The neural basis of timing: Distributed mechanisms for diverse functions. *Neuron, 98*(4), 687–705. https://doi.org/10.1016/j.neuron.2018.03.045

Pavailler, N., Gevers, W., & Burle, B. (2024). Temporal metacognition: direct readout or mental construct? the case of introspective reaction time. *Journal of Experimental Psychology: General*.

Pereira, M., Perrin, D., & Faivre, N. (2022). A leaky evidence accumulation process for perceptual experience. *Trends in Cognitive Sciences, 26*(6), 451–461.

Sanchez, R., Courant, A., Desantis, A., & Gajdos, T. (2024). Making precise movements increases confidence in perceptual decisions. *Cognition, 249*, Article 105832. https://doi.org/10.1016/j.cognition.2024.105832

Wang, J., Narain, D., Hosseini, E. A., & Jazayeri, M. (2018). Flexible timing by temporal scaling of cortical responses. *Nature Neuroscience, 21*(1), 102–110. https://doi.org/10.1038/s41593-017-0028-6

Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics, 14*(1), 5–12. https://doi.org/10.3758/BF03198607

Yallak, E., & Balcı, F. (2021). Metric error monitoring: Another generalized mechanism for magnitude representations? *Cognition, 210*, Article 104532. https://doi.org/10.1016/j.cognition.2020.104532

Van Wassenhove, V. (2009). Minding time in an amodal representational space. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1525), 1815–1830.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1310–1321.

5.3) Paper 3 : *The Brain as Actor and Evaluator: Distinct Neural Codes for Timing and Self-Evaluation of Timing Errors Revealed by Transformer-Based Decoding*

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

5.3) Paper 3

Title: The Brain as Actor and Evaluator: Distinct Neural Codes for Timing and Self-Evaluation of Timing Errors Revealed by Transformer-Based Decoding

Authors: Sena N. Bilgin[1]*, Dunia Giomo[1,3], Urfan Mustafali[4] , Tadeusz W. Kononowicz[1,2]*

Corresponding author(s): Tadeusz W. Kononowicz (tadeusz.kononowicz@cnrs.fr), Sena N. Bilgin (sbilgin@sd.psych.pan.pl)

Affiliations:

[1] Institute of Psychology, Polish Academy of Sciences, ul. Jaracza 1, 00-378 Warsaw, Poland

[2] Paris-Saclay University, CNRS, Paris-Saclay Institute of Neuroscience (NeuroPSI), 91400 Saclay, France

[3] Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy

[4] School of Human Sciences, VIZJA University, ul. Okopowa 59, 01-043 Warsaw, Poland

Author Contributions: SNB, TWK, and DG conceived and designed the study. DG and UM conducted EEG experiments with human participants. SNB developed the EEG denoising and artificial intelligence (AI) pipelines, preprocessed the data, and performed the AI-based modelling. SNB and TWK planned and executed the analytical approach. DG contributed to the interpretation of the results. SNB drafted the initial manuscript; all authors critically reviewed it, provided feedback, and contributed to refining the research, analyses, and manuscript.

Acknowledgements: This research was supported by the National Science Centre, Poland (grant no. 2019/35/B/HS6/04389 to TWK). We thank Hatice Balkan for assistance with EEG data collection. We also thank Izabela Szumska and Dorota Stelmaszyńska for technical assistance in the experimental setup.

Competing Interest Statement: No conflict of interest.

Keywords: EEG-decoding, error monitoring, metacognition, interval timing, time production, vision transformers (ViT), multi-head attention, principal component analysis.

This PDF contains the Main Text, Figures 1–5, and Table 1.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Metacognitive self-evaluation, the capacity to assess one's own performance without external feedback, is fundamental to adaptive cognition. A central unresolved question in metacognition is whether self-evaluation reflects a direct readout of the neural signals supporting first-order performance, or whether it depends on partially distinct higher-order representations. Unusual approach of monitoring actions in a continuous dimension, such as time, is particularly useful because errors are graded rather than binary, allowing metacognition to be treated as a quantitative estimate of error magnitude. Temporal error monitoring (TEM), in which individuals judge the accuracy of self-generated time intervals using only internal signals, provides a tractable framework for isolating metacognitive evaluation from external correction. We asked whether TEM engages the same neural mechanisms as first-order timing performance, the execution and control of the timed action itself, or recruits at least partially distinct processes involving additional evaluative computation. We decode single-trial EEG using a Vision Transformer applied to PCA-optimized θ, α, and β signals recorded during both monitoring of motor timing. First-order timing was decodable from each frequency band independently, whereas TEM required simultaneous integration across all three bands. Individuals for whom TEM neural states were decoded with greater accuracy demonstrated stronger behavioral error-monitoring precision. These findings establish TEM as a neurally distinct re-representation of timing performance, demonstrating that metacognitive self-evaluation requires evaluative computations beyond those supporting motor action execution.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1) Introduction

The capacity to monitor and evaluate one's own cognitive processes, metacognition, is fundamental to adaptive behaviours across species.[1,2] Temporal error monitoring (TEM) is one well-defined instance of this capacity: the ability to assess the accuracy of self-generated time intervals without external feedback. By isolating self-evaluation from exogenous cues, TEM offers a powerful framework for probing the neural basis of self-monitoring. A central question, however, remains unresolved: do metacognitive evaluations arise from the same decision variable as first-order performance, or do they at least partially engage distinct processes supported by additional neural representations?[3,4]

Timing tasks are particularly well-suited to address this question: unlike most cognitive domains, they require reliance exclusively on internal representations, with no external cues available. In motor timing paradigms, an agent generates a time interval by initiating it with one action and terminating it with a second when the target duration is judged to have elapsed. The discrepancy between intended and produced duration constitutes the error signal; estimating its magnitude and direction in the absence of feedback is the core requirement of TEM. Reliable TEM performance across both humans[5-10] and rats[11]confirms that internally generated signals alone are sufficient to sustain metacognitive evaluation.

Accounts differ, however, in what they propose about the architecture that supports it. Single-process accounts hold that confidence and performance share a common decision variable, such that metacognitive judgements directly reflect task-execution signals with no additional processing required.[12] Hierarchical models, such as the Bayesian framework of Fleming and Daw,[13] posit a functionally distinct metacognitive stage; yet this stage still derives its content directly from first-order signals, computing confidence from their strength or uncertainty, making metacognition temporally later but not neurally distinct. Higher-Order Representation (HOR) theory draws a sharper distinction: conscious awareness of a mental state requires a higher-order representation that takes the first-order state as its object, a genuine re-representation rather than a downstream derivation from first-order signals.[14] Neural evidence supports this view, implicating specialized prefrontal circuits in higher-order representations that cannot be reduced to first-order activity.[3,15]

Each of these positions generates a distinct prediction for TEM. Drift-diffusion models hold that error evaluation arises from the same accumulator governing interval production, implying shared neural substrates and a tight trial-by-trial coupling between timing performance and TEM precision.[12,16] Temporal Metacognition, by contrast, views error

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

evaluation as an active inferential process in which the brain decodes and re-represents internal duration signals at a higher level, not a mere reflection of the timing signal but a higher-order operation that draws on, yet goes beyond, first-order timing information, consistent with HOR theory.[4,14]

Distinguishing between these predictions, however, requires evidence that condition-averaged analyses cannot provide. Prior work linked beta-band oscillations to timing accuracy[17,18] and metacognitive sensitivity,[4] alpha oscillations to clock-speed modulation[19] and temporal expectation,[20] and frontal midline theta to prediction errors scaling with error magnitude.[21] Whether these mechanisms are shared between first-order timing and metacognitive evaluation, or whether metacognition requires their coordinated interaction, cannot be resolved from averaged data alone.

To address this, we applied a Vision Transformer decoding architecture[22,23] to single-trial EEG recorded during a motor timing task with no external feedback. First-order timing decodability did not predict trial-by-trial TEM precision, challenging drift-diffusion accounts,[12,16] while metacognitive state decodability predicted TEM precision independently of timing performance. These findings argue against single-process and hierarchical accounts and support HOR theory, establishing metacognitive evaluation as an at least partially distinct re-representation of the brain's self-evaluative architecture.

2) Methods

2.1) Subjects

A total of 29 participants (17 male, 12 female; age range: 18–36 years; M = 24.10, SD = 4.28) were recruited. All participants reported normal or corrected-to-normal vision. Written informed consent was obtained from each participant before data collection. All procedures were approved by the Research Ethics Committee at the Institute of Psychology of the Polish Academy of Sciences (reference: IP.403.3.2022) and conducted in accordance with the Declaration of Helsinki.[24]

2.2) Apparatus

Stimulus presentation. Subjects sat at approximately 55–60 cm distance from the computer monitor. Visual stimuli, task instructions, and feedback were presented using the

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Psychophysics Toolbox[25] implemented in MATLAB. Behavioral responses were recorded via keyboard and mouse.

EEG recording. Electroencephalographic data were acquired using a 64-channel actiCAP snap cap (Easycap) with gel-based electrodes and an ActiCHamp amplifier (Brain Products). The EEG signal was sampled at 500 Hz. Electrode impedances were maintained below 10 kΩ throughout each recording session.

Data processing and analysis. All preprocessing and analysis pipelines were implemented in Python (v3.12.5). EEG signal processing was performed using MNE-Python (v1.9.0), and deep learning models were built in PyTorch (v2.5.1) within the TorchEEG framework. Code is available as described in the *Code Availability* statement.

2.3) EEG Experiment Procedure

This study analyzed a subset of data (approximately 25%) from a larger study combining EEG and transcranial direct current stimulation (tDCS). EEG was recorded from a custom 64-channel cap with openings at Fpz and Fz, where neurostimulation electrodes were placed (anode at Fpz, cathode at Fz). Data were therefore recorded from 63 channels, with those two sites excluded.

The full experiment included a tDCS session and a Sham session, conducted on separate days to reduce cognitive fatigue and potential tDCS aftereffects (see Supp. Mat. Table 1). Each session comprised two blocks of approximately 20 minutes each. The first block was either a tDCS or Sham block, with order counterbalanced across participants. During the tDCS block, stimulation was delivered continuously at 1 mA, whereas in the Sham block, current was applied only briefly at onset before being discontinued to mimic the sensation of stimulation.

The second block consisted of an EEG-only recording. Because tDCS can induce aftereffects that persist beyond the stimulation period, only the EEG-only block of the Sham session yields uncontaminated measures of brain activity. Accordingly, all analyses in this study were restricted to this block. A 5-minute break was provided between blocks to minimize fatigue. Event markers were transmitted from MATLAB to the EEG system to label critical time points within each trial, enabling the extraction of epochs corresponding to pre-event, event, and post-event periods relevant to TEM.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Within this EEG-only recording, participants performed both time production (TP) and TEM tasks (Figure 1). Specifically, they were required to reproduce a 2-second target duration using a computer mouse. The target duration was introduced and practiced during a structured training phase to ensure that participants developed a stable internal representation of the interval.

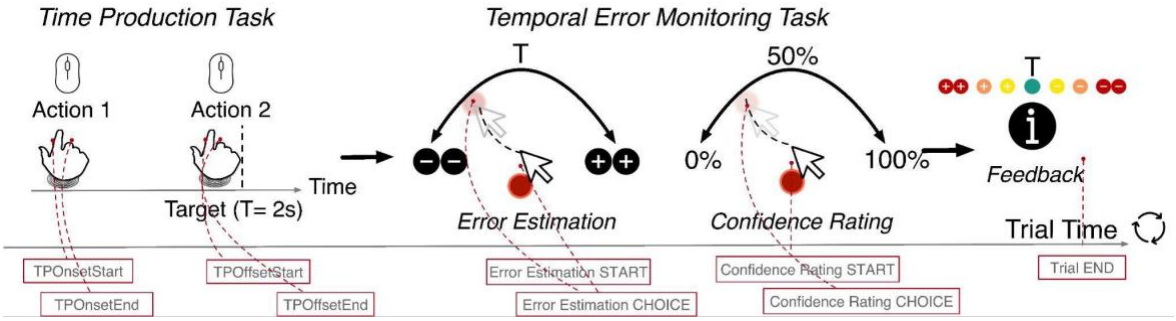


Figure 1. Experimental procedure. Participants reproduced a 2-second target interval by pressing a mouse button to mark its onset and offset. Following each reproduction, they provided two metacognitive judgments: a temporal error estimate indicating the perceived direction and magnitude of their timing error, and a confidence rating reflecting certainty in that estimate. Behavioral measures extracted from each trial included produced duration, temporal error estimation, and TEM confidence. The Metacognitive Inference (MI) index quantified trial-by-trial correspondence between objective timing performance and subjective error reports. Together, these measures enabled independent assessment of timing precision, metacognitive sensitivity, and confidence calibration.

a) Training Phase

Training was structured into three successive phases; each designed to progressively establish and refine the components of the task.

In the first phase, participants were familiarized with the 2-second target duration through 10 visually guided trials. Each trial began with a white fixation cross, which turned green to signal the onset of the interval; after 2 seconds, the green cross disappeared, followed by a randomly selected inter-trial interval (ITI) of 1–2 seconds. Participants were instructed to internalize the duration without relying on counting strategies. During the first session, this phase was repeated for an additional 10 trials if the target interval had not yet been reliably

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

memorized. In the second session, all participants completed a full set of 10 trials to ensure consistency.

Building on this internal representation, the second phase focused on time production (TP). Participants completed 30 trials in which they reproduced the target duration by using a computer mouse to mark both the onset and offset of the interval. Each trial began with a white fixation cross indicating the start of the response. Following each attempt, visual feedback informed participants whether their reproduction fell within ±1000 ms of the target; trials exceeding this threshold were classified as invalid and excluded from further analysis. This TP training phase was not repeated in the second session.

Finally, the third phase introduced metacognitive evaluation through temporal error monitoring (TEM). In 10 trials, participants first reproduced the target duration and then provided two judgements: (i) an estimate of the direction and magnitude of their timing error on a continuous scale, and (ii) a confidence rating for that estimate. Both responses were collected using a custom arc-based interface (Figure 1), in which participants dragged the mouse from the center toward a position on the arc, allowing for high-resolution and spatially unbiased input.

To facilitate calibration of these metacognitive judgements, participants received explicit feedback on the magnitude of their timing errors during this phase. Feedback was individualized based on each participant's performance variability, quantified as the standard deviation (SD) of TP errors obtained in the second phase. Errors were categorized into four levels: *too short/long* (>2 SD), *quite short/long* (>1 SD), *a little bit short/long* (>0.3 SD), and *very close to the target* (≤0.3 SD). On Day 2, participants completed an additional 15 TEM training trials, with feedback calibrated using the TP standard deviation derived from Day 1.

b) Main Experimental Phase

During the main phase, each trial began with participants initiating the interval by pressing a button and terminating it with a second press once they judged that 2 seconds had elapsed, constituting the TP. Following this, they provided two metacognitive judgements. The TEM phase was separated from the TP response by a brief 300 ms interval, measured from the offset of the produced interval to the onset of the judgement phase.

Feedback availability was manipulated within the session: 50% of trials included feedback, while the remaining 50% did not. On feedback-present trials, participants received performance feedback based on the same individualized deviation categories established during TEM training. In contrast, on feedback-absent trials, no external information was

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

provided, requiring participants to rely exclusively on their internal representations of time to evaluate their performance. This manipulation enabled a direct assessment of TEM under conditions with and without external feedback.

Computation of TEM variables and confidence ratings. Both judgements were recorded via the arc-based interface introduced during training. The arc comprised 101 evenly spaced points spanning -45 ° to +45 °. For each mouse click, the arc coordinate with the smallest Euclidean distance to the click position was selected as the raw response. This response was then mapped onto a standardized continuous scale. Temporal error estimates were normalized from −100% to +100%, where −100% indicated maximum underestimation and +100% maximum overestimation. Confidence ratings were normalized from 0% to 100%. This ensured both measures were represented on continuous, trial-specific scales suitable for single-trial analysis.

Behavioral metrics extracted from each trial included TP duration, temporal error response, and TEM confidence rating. A Metacognitive Distance (MD) index was computed as the absolute difference between the z-scored temporal error response and the z-scored TP duration, providing a trial-wise measure of alignment between temporal error estimates and actual timing performance. Formally, for trial (i):

$$MD_i = |z(temporal\ error\ response_i) - z(TP\ duration_i)|$$

TP categorization. Continuous TP values were discretized into three groups using the 33rd and 66th percentiles of the TP distribution as thresholds. Trials below the lower threshold were labelled “Short”, those above the upper threshold were labelled “Long”, and trials between the two thresholds were labelled “Correct”. This tertile-based categorization yielded an approximately balanced distribution across the three classes and served as the multiclass classification target.

MI categorization. MD values were dichotomized using a median split across all trials. Trials below the median were classified as “High MI” and those at or above as “Low MI”, on the assumption that smaller absolute differences reflect more accurate metacognitive inference. This yielded approximately equal trial counts across groups and defined the binary classification target.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

TEM confidence categorization. TEM confidence values were similarly dichotomized using a median split. Trials below the median were assigned to the "Low Confidence" group and those at or above to the "High Confidence" group, providing a balanced binary classification target for subsequent analyses.

Overall, these categorizations enabled evaluation of motor action timing performance and metacognitive precision at the single-trial level.

3) Statistical Analysis

Behavioral data were analyzed in Python (version 3.12) using the *statsmodels* package[26] (version 0.14.5). Statistical analyses included ordinary least squares regression, analysis of variance, and pairwise comparisons. Single-trial measures of TP and TEM were preprocessed by excluding outliers exceeding 3.5 standard deviations from each participant's mean.

Relationships among continuous variables were assessed using linear mixed-effects models (LMMs) estimated via restricted maximum likelihood (REML). Trial-level observations were nested within participants to account for repeated measures. Fixed effects captured population-level relationships, whereas random effects modeled between-participant variability. Model selection was guided by the Akaike Information Criterion (AIC) and likelihood ratio tests for nested comparisons. Model-derived slopes were visualized with 95% confidence intervals.

Group-level differences in decoding accuracy and behavioral measures were evaluated using independent-samples *t*-tests. This analytical framework enabled robust estimation of effects while appropriately accounting for repeated measures and inter-individual variability.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4) EEG Data Processing

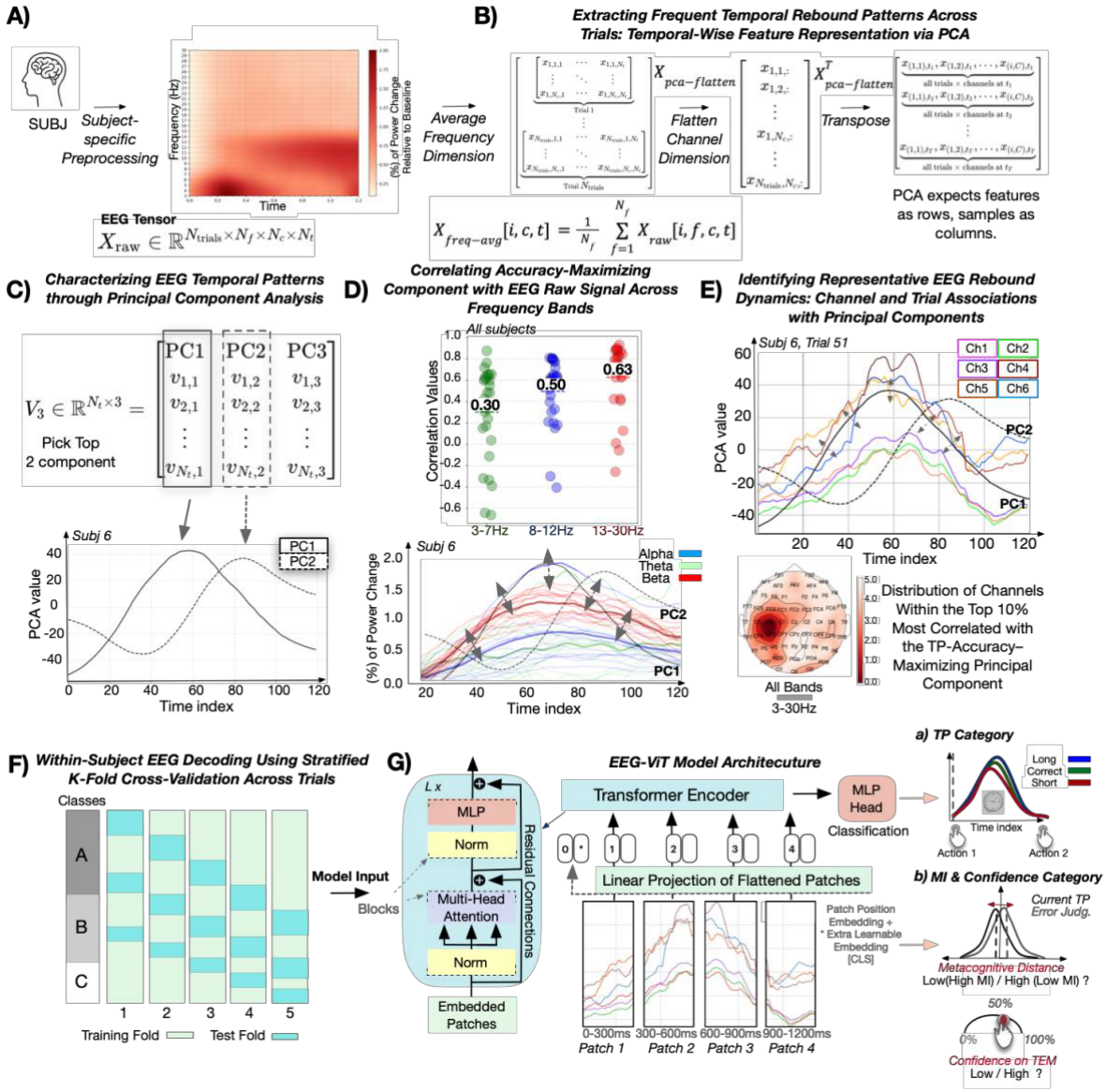


Figure 2. Subject-specific preprocessing pipeline and Vision Transformer architecture for EEG-based behavioral decoding.

a EEG preprocessing and data structuring. Raw EEG underwent artifact rejection, bandpass filtering (1–120 Hz), notch filtering (50 Hz), and baseline normalization. Preprocessed signals were structured as four-dimensional tensors (trials × channels × frequencies × time). The right panel shows a time-frequency representation averaged across

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

participants and trials, with color indicating percentage power change relative to baseline (−0.3 to −0.1 s before TP onset). See Methods 3.1–3.2.

b Data preparation for PCA. EEG signals were frequency-averaged across three bands (θ: 4–7 Hz; α: 8–12 Hz; β: 13–30 Hz) and flattened across channels and trials to reduce dimensionality and extract dominant temporal patterns. The resulting matrix was transposed into a trials × channels × time representation suitable for principal component analysis (PCA), with each trial treated as a feature to identify common temporal patterns across channels. See Methods 3.2a–d.

c Principal component extraction. PCA was applied to the covariance matrix of the transposed data. Eigenvalue decomposition yielded temporal components ranked by explained variance. The first two components (PC1 and PC2) were retained to capture the dominant temporal structure across trials. Example waveforms for participant 6 are shown in the bottom panel. Post-movement β activity showed considerable variability, with no standardized temporal range defining a canonical β rebound. The accuracy-maximizing component was defined as the one yielding the highest decoding performance. Similar rebound patterns were observed in the α and θ bands. See Methods 3.2e–g.

d Validation of extracted components against physiological oscillations. To assess whether PCA-derived components reflected physiologically meaningful activity, the accuracy-maximizing component was correlated with the raw EEG signal in each frequency band. β-band activity showed the strongest associations ($M \approx 0.63$), followed by α ($M \approx 0.50$) and θ ($M \approx 0.30$). Pairwise comparisons using Tukey's HSD confirmed that β correlations were significantly higher than θ ($MD = -0.33$, $p = .007$, 95% CI [−0.58, −0.08]), while α did not differ significantly from β ($MD = 0.14$, $p = .396$) or θ ($MD = -0.19$, $p = .172$). These findings confirm that the accuracy-maximizing component most strongly captured β-rebound dynamics, with α and θ contributing more modestly. The bottom panel illustrates this component for participant 6, with frequency-specific correlations highlighted in red (β), blue (α), and green (θ).

e Selection of informative channels and trials. Correlations between the accuracy-maximizing component and frequency-averaged EEG signals were computed to identify the most informative signals for model input. Channels were ranked first, followed by trials, with the top 10% of channels and top 30% of trials retained per participant. Example data from

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

participant 6, trial 51, are shown in the top panel, with each waveform representing one of the six top-ranked channels. The bottom topographic plot shows channels consistently selected across participants, with clusters over central and parietal regions overlapping with known post-movement β rebound areas, supporting the functional relevance of the selection procedure. See Methods 3.2h.

f Stratified cross-validation design. Within-participant decoding used stratified 5-fold cross-validation, maintaining proportional representation of behavioral categories (TP, MI, and TEM confidence) across folds. This minimized class-imbalance bias and prevented information leakage between training and test sets. See Methods 4d.

g Vision Transformer architecture for temporal pattern learning. Single-trial EEG signals (trials × channels) were segmented into contiguous temporal patches, each linearly projected into fixed-dimensional embeddings. A learnable class token ([CLS]) was prepended to the patch sequence to aggregate trial-level information, with its final state serving as the classification input, following established NLP architectures such as BERT[27]. Positional encodings preserved temporal order across the sequence. Embeddings were passed through six stacked transformer encoder layers, each containing multi-head self-attention to identify informative temporal relationships, layer normalization for training stability, residual connections for gradient flow, and feedforward MLP blocks. The final [CLS] embedding was classified via MLP to predict behavioral categories (TP, MI, and TEM confidence). This architecture captured both local and global temporal dynamics, enabling identification of distributed EEG patterns that predict trial-level behaviours. See Methods 4a–c.

4.1) Signal Denoising

Raw EEG data were processed using MNE-Python[28,29] (v1.9.0). Data were loaded from BrainVision files, and irrelevant segments, such as inter-block breaks, were removed. One participant (nr. 18) was excluded due to excessive α-wave activity, yielding a final sample of 28 participants.

EEG signals were initially re-referenced to the average of all channels to eliminate global shifts. Poor-quality channels were identified and excluded before Independent Component Analysis (ICA). A 50 Hz notch filter was applied to remove power-line interference, and a bandpass filter (1–120 Hz) was applied to remove low-frequency drift and high-frequency

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

noise. ICA decomposed the signal into independent components, which were visually inspected to identify and remove artifact-related components, including eye movements (EOG), muscle contractions (EMG), and heartbeats (ECG). Bad channels were subsequently interpolated, and data were re-referenced to the channel average.

Artifact rejection employed the AutoReject protocol[30] (v0.4.2), with interpolation levels (1, 4, 16, 32) and consensus percentages (0–1.0) optimized via Bayesian methods. Time-frequency analysis was performed using multitaper decomposition across θ (3–7 Hz), α (8–12 Hz), and β (13–30 Hz) bands, as well as a merged band combining all three. Cycle parameters for each band were selected to balance time and frequency resolution, reducing spectral leakage and improving frequency estimation precision.

Baseline correction was applied relative to a reference window of −0.3 to −0.1 s before the first keypress. Time-frequency data were then cropped from TP onset to 1.2 s post-onset, capturing frequency dynamics during the TP phase. The TP offset marker was excluded to ensure that model predictions relied solely on ongoing EEG activity rather than the interval termination.

4.2) Signal Filtering for Model Input

a) Raw EEG Data Structure:

EEG data were preprocessed and analyzed at the individual-participant level to account for inter-subject variability in noise characteristics and neural activity patterns.[31] This participant-specific approach mitigates a well-established limitation of group-level EEG analyses, in which pooling data across participants can obscure fine-grained, trial-level neural dynamics relevant to the decoding of cognitive states.

Following denoising and artefact correction, clean EEG signals were organized into a four-dimensional tensor:

$$X_{raw} \epsilon\ R^{N_i \times\ N_f\ \times N_c \times N_t}$$

where N_i denotes the number of trials, N_f the number of frequency bins derived from time-frequency decomposition, N_c the number of EEG channels, and N_t the number of time points capturing within-trial temporal dynamics (Figure 2a). This representation preserves the joint spectral, spatial, and temporal structure of the EEG signal, which is frequently collapsed or averaged in conventional preprocessing pipelines.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For TP, MI, and TEM confidence classification based on EEG activity during the TP phase, each trial comprised 120 time points spanning 0–1.2 s relative to the event marker. For MI classification based on activity during the temporal error response and TEM confidence judgement phases, each trial comprised 60 time points spanning 0–0.6 s, reflecting the shorter temporal scale of post-response evaluative processing.

b) Frequency-Averaged Data Representation:

EEG data were averaged across the frequency dimension to isolate temporally structured neural dynamics independently of band-specific spectral contributions, producing a three-dimensional representation that preserved trial-wise spatial and temporal information (Figure 2b):

$$X_{freq-avg}\,[\,i\,,c\,,t\,] = \frac{1}{N_f}\ \sum_{f=1}^{N_f}\quad X_{raw}[\,i,f,c,t\,]$$

resulting in:

$$X_{freq-avg} \in\ R^{N_i \times\ N_c\ \times N_t}$$

Each trial-specific slice represented the multichannel EEG time series collapsed across frequencies, with channels along rows and time points along columns. Collapsing across frequencies reduced input dimensionality and mitigated redundancy from correlated spectral components, while preserving the temporal dynamics of neural activity. This representation enabled a direct test of whether temporally organized activity alone was sufficient to decode task-relevant cognitive and metacognitive states. Model performance on frequency-resolved versus frequency-averaged inputs was then compared to assess the relative contribution of oscillatory specificity beyond broadband temporal structure. This contrast clarified the extent to which trial-by-trial variability in motor timing, error monitoring, and TEM confidence was driven by temporal patterns rather than frequency-specific neural dynamics.

c) Flattened Trial-Channel Matrix for PCA:

To characterize dominant temporal modes of variability shared across trials and recording sites, the frequency-averaged EEG data were reshaped into a two-dimensional matrix suitable for PCA. The trial and channel dimensions were collapsed into a single observation axis:

$$X_{pca-flatten} \in\ R^{(N_i\ .\ N_c) \times N_t}$$

where each row corresponded to an individual trial–channel observation and each column to a time point (Figure 2b). Treating each trial–channel pair as a distinct realization of the underlying temporal dynamics preserved the full temporal structure along the feature axis. This

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

formulation directed PCA toward low-dimensional temporal components capturing shared variance across trials and channels, rather than idiosyncratic spatial or trial-specific fluctuations.

d) Transposed Matrix for PCA:

To extract shared temporal structure, the flattened matrix was transposed such that time points served as variables and trial–channel pairs as observations:

$$X^{T}_{pca-flatten} \in R^{N_t \times (N_i \cdot N_c)}$$

Each row thus represented the percentage change in EEG power relative to baseline across all trial–channel observations at a given time point (Figure 2b). This orientation directed PCA towards capturing temporal correlations across trials and channels, which enabled the identification of common temporal patterns.

e) Covariance Matrix:

After mean-centering the transposed matrix across observations,

$$\underline{X}^{T} = X^{T}_{pca-flatten} - mean\,(X^{T}_{pca-flatten}, axis = 1)$$

the temporal covariance matrix was computed as:

$$\Sigma = \frac{1}{N_i \cdot N_c - 1}\ \underline{X}^{T}\ (\underline{X}^{T})^{T} \in R^{N_t \times N_t}$$

Each element Σ_{t_1, t_2} quantified the covariance between time points t_1 and t_2 across all trial–channel observations. Diagonal elements reflected temporal variance, while off-diagonal elements captured temporal correlations. This covariance structure formed the basis for extracting principal temporal components.

f) Principal Components:

Eigen-decomposition of the covariance matrix Σ yielded eigenvectors V and eigenvalues Λ:

$$\Sigma V = V\Lambda$$

where $V \in R^{N_t \times N_t}$ contained eigenvectors as columns and Λ was the corresponding diagonal eigenvalue matrix. The first three eigenvectors,

$$V_3 \in R^{N_t \times 3}$$

represented the dominant temporal activity patterns, with each column corresponding to a principal component (PC1, PC2, PC3) and each row to a time point (Figure 2c). These components summarized the primary temporal structure shared across trials and channels, accounting for the largest proportion of total variance.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

g) PCA Scores:

Each trial–channel observation was projected onto the dominant temporal components by multiplying the flattened matrix by the top eigenvectors:

$$Z = X_{pca-flatten} V_3 \in R^{(N_i . N_c) \times 3}$$

Each row Z represented a trial–channel observation, and each column its loading on one of the three principal temporal components. This produced a compact representation of the EEG data while retaining both trial- and channel-level information.

h) Channel and Trial Selection Based on Principal Components' Correlations:

EEG trials and channels vary in how strongly they express dominant temporal patterns, reflecting differences in neural responses, cognitive state, and noise. The first two principal components (PC1 and PC2, collectively V_2) were extracted from the frequency-averaged EEG data to isolate the most informative and reproducible signals. PC1 and PC2 captured dominant temporal dynamics consistent with β-band rebound activity, previously implicated in motor action timing[17,18] and temporal error monitoring.[4]

For each component V_k (k = 1, 2), the average correlation between the component and each channel's EEG signal across all trials was computed:

$$r_{c,k} = \frac{1}{N_i} \sum_{i=1}^{N_i} corr\,(X_{freq-avg}\,[\,i, c, :\,]\,, V_k\,)$$

Channels were ranked by their average correlations with PC1 and PC2, and the top 10% (six channels) were retained for further analysis (Figure 2e):

$$X_{chan-filtered}\,[\,i, c, t\,] = X_{freq-avg}[\,i, C_{top}\,, t\,]$$

For each trial i, the retained channel signals were averaged:

$$\underline{X}_{top-ch}\;\;[\,i\,, t\,] = \frac{1}{N_{top}} \sum_{c \in C_{top}} X_{chan-filtered}[\,i, c, t\,]$$

Trials were subsequently ranked by their correlation with PC1 and PC2:

$$r_{i,k} = corr\,(\,\underline{X}_{top-ch}\;\;[\,i, :\,]\,, V_k(t)\,)$$

This ranking prioritized trials most strongly expressing the shared temporal dynamics of interest, reducing the influence of noisy or weakly aligned observations. Trial subsets corresponding to the top 30%, 50%, and 100% were retained to evaluate how alignment with

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

dominant temporal patterns affects downstream model performance. The resulting trial–channel–time representations, restricted to selected channels and ranked trial subsets,

$$X_{ch\,,trial-filtered}\,[\,i,c,t\,]\,=X_{ch-filtered}\,[\,I_{top},C_{top},t\,]\;,\quad i\,\epsilon\,I_{top}\;,\;\;c\,\epsilon\,C_{top}\;,\;t=1,\ldots N_t$$

served as inputs for all subsequent predictive analyses. Comparing model performance across trial subsets showed that prioritizing channels and trials aligned with shared temporal dynamics improved decoding accuracy and the generalizability of learned neural representations.

5) EEG Vision Transformer (EEG-ViT) Model Pipeline

a) Model Input Representation:

Following channel and trial selection, the preprocessed EEG data were organized as:

$$X_{ch,trial-filtered}\,[\,i,c,t\,]\,\epsilon\,R^{(N_i\times N_c\times N_t)}$$

This filtered dataset retained only the most informative trials and channels, improving signal-to-noise ratio and focusing downstream modelling on dominant temporal patterns. Each trial–channel pair was treated as an independent observation, and the 3D tensor was reshaped into a 2D matrix:

$$X_{input}\,\epsilon\,R^{(N_{trials}\cdot\,N_c)\,\times\,N_t}$$

Each row corresponded to the temporal EEG signal from a single trial–channel pair, and each column to a time point. This reshaping emphasized temporal structure while allowing patterns to be learned jointly across trials and channels, without imposing a fixed spatial topology.

b) Patch Embedding and Input Construction:

Transient, band-limited power rebounds, such as β-band rebounds, are hallmark neural signatures of motor action timing and temporal error monitoring. These events are brief, variable in latency and amplitude across trials and channels, and embedded within substantial background noise. Their reliable detection requires two complementary capacities: *(i)* local representations capturing short temporal segments of power modulation, and *(ii)* a global integration mechanism that combines local signals across the trial to form a coherent representation of the underlying neural state.

To satisfy both requirements, the one-dimensional EEG time series was treated as *a temporal image* and modelled using a Vision Transformer (ViT) architecture adapted for single-trial EEG,[22] with our minor modifications to accommodate trial-specific temporal dynamics (Figure 2g). Temporal patch segmentation is conceptually analogous to spatial patch decomposition in image ViTs: each patch encodes local temporal (and implicitly spectral)

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

structure within a short window, while multi-head self-attention captures interactions between patches across the full trial. Critically, unlike convolutional filters, which impose fixed local receptive fields, self-attention can model dependencies across arbitrary temporal distances, enabling the detection of events whose diagnostic relevance emerges from broader temporal context (e.g., a rebound following earlier suppression).

The ViT was configured with a grid size of (1,1) and a spatial patch size of (1,1), treating each channel as a single spatial unit. Each temporal sequence was divided into $N = T/P_t = 4$ non-overlapping patches of length P_t = 30 time points (see Table 1 for full hyperparameters). For sample *i,* patch *j* was defined as:

$$p_{i\,j} = X_{input,i}\,[\,(j-1)\,P_t{:}\,j\,P_t\,) \in R^{P_t}$$

Each patch was linearly projected into a 64-dimensional embedding space (d = 64) via a learnable mapping:

$$z_{i,j} = W_{patch}\,p_{i,j} + b_{patch} \in R^{d} \quad , \quad W_{patch} \in R^{d \times P_t}$$

A learnable classification token $[CLS] \in R^{d}$ was prepended to the patch sequence, and learnable positional encodings $pos \in R^{(N+1) \times d}$ were added to preserve temporal order:

$$Z_i^{(0)} = [\,[CLS]\,;\, z_{i,1}, z_{i,2}, \ldots, z_{i,N}] + pos, \qquad Z_i^{(0)} \in R^{(N+1) \times d}$$

Dropout (p = 0.15) was applied to the embedded sequence to improve regularization.

c) Transformer Encoder:

The transformer encoder comprised L= 6 identical layers, each consisting of a multi-head self-attention (MHA) block followed by a position-wise feed-forward network (MLP), with residual connections and layer normalisation applied at each stage.

Multi-head Self-Attention (MHA):

For a layer l, the input $X^{(l-1)} \in R^{B \times (N+1) \times d}$ was normalized:

$$U^{(l)} = LayerNorm(X^{(l-1)})$$

For each head h, queries (Q), keys (K), and values (V) were obtained via linear projections:

$$Q = U^{(l)}\,W_h^Q\,, \quad K = U^{(l)}\,W_h^K\,, \quad V = U^{(l)}\,W_h^V, \quad W_h^Q,\ W_h^V\,,\ W_h^V \in R^{d \times d}$$

Attention was computed independently across $H = 6$ heads, each with a per-head dimension $d_h = 32$:

$$Attn_h(\,Q_h, K_h, V_h\,) = softmax\,(\frac{Q_h {K_h}^T}{\sqrt{d_h}})V_h$$

Head outputs (W^o) were concatenated and projected:

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$$MHA\ (U^{(l)}) = Concat(Attn_1, \ldots, Attn_H)\ W^o, \qquad W^o \in R^{d \times d}$$

with residual update:

$$X'^{(l)} = X^{(l-1)} + MHA\ (U^{(l)})$$

Feed-Forward (MLP) Block:

The normalized intermediate representation,

$$V^{(l)} = LayerNorm(X'^{(l)})$$

was passed through a position-wise MLP with a hidden dimension $d_{ff} = 64$:

$$MLP\ (V^{(l)}) = W_2(W_1 V^{(l)} + b_1) + b_2$$

with residual update:

$$X^{(l)} = X'^{(l)} + MLP(V^{(l)})$$

This process was repeated for $l = 1, \ldots, L$.

Pooling and Classification:

Following the final encoder layer, the classification token embedding was extracted:

$$h_i = Z_i^{(L)}[\,0\,, :\,] \in R^d$$

This vector provided a compact summary of the temporally distributed EEG dynamics for each trial–channel pair and was passed to a linear classifier to produce task-specific predictions:

$$\hat{y}_i = W_{cls}\, h_i + b_{cls}\ , \qquad \hat{y}_i \in R^C\ , \quad C = 3$$

corresponding to the target label categories (e.g., short, correct, long for TP; low versus high for TEM confidence) (Figure 2g, insets a–c).

d) Model Fitting:

A stratified k-fold cross-validation (k = 5) was applied at the trial level to ensure balanced representation of both dataset size and target categories across folds (Figure 2f). In each iteration, k−1 folds were designated for training, and the remaining fold for testing, with the model reinitialized before each fold to prevent leakage of optimized parameters. Constructing folds at the trial level ensured that no trial in the test set appeared in the training set, eliminating artificial inflation of performance due to correlations across channels within the same trial. Stratification further mitigated class imbalance, yielding stable and unbiased evaluation metrics[32].

Batches lacking examples from all target categories were skipped to prevent gradient updates from favoring overrepresented classes. This cross-validation framework maximized

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

data utilization, balanced the bias–variance trade-off, and produced reproducible estimates of model performance across trials and participants. Evaluating accuracy across independent folds confirmed the generalizability of temporally structured neural representations, supporting applicability to diverse datasets, experimental paradigms, and cognitive processes[33].

Parameter	TPOnsetStart Event	TEMOnsetStart Event
Input chunk size	120 time points	60 time points
Temporal patch size (Pt)	30	15
Number of patches (N)	4	4
Embedding dimension (d)	64	64
Transformer depth (L)	6 layers	6 layers
Attention heads (H)	6	6
Head dimension (dh)	32	32
Feed-forward channels (dff)	64	64
Dropout rate	0.15	0.15
Embedding dropout	0.15	0.15
Grid size	(1, 1)	(1, 1)
Spatial patch size	(1, 1)	(1, 1)
Output classes	TP (Short, Correct, Long); MI (Low, High); Confidence (Low, High)	MI (Low, High); Confidence (Low, High)
Epochs	120	120
Batch size	18	18
Optimizer	AdamW (lr=5e-4, weight_decay=1e-2)	AdamW (lr=5e-4, weight_decay=1e-2)
Learning rate scheduler	CosineAnnealingLR (Tmax=50, eta_min=1e-3)	CosineAnnealingLR (Tmax=50, eta_min=1e-3)

Table 1. Hyperparameters for the ViT-based EEG decoding pipeline. The table depicts the hyperparameters and selected values used in the ViT-based EEG decoding pipeline for patch embedding, transformer architecture, and model training, together with brief notes on parameter ranges.

6) Results

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

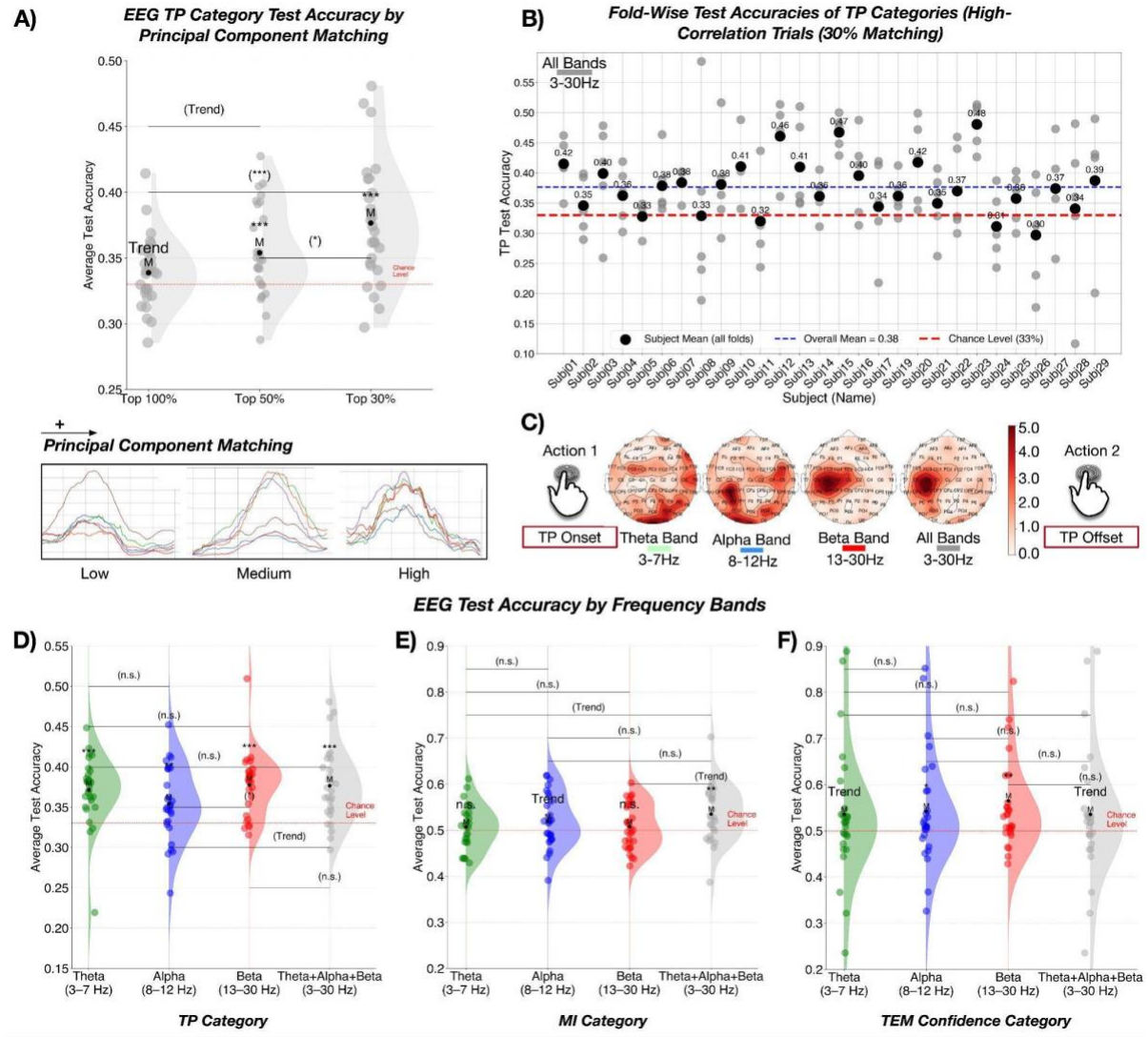


Figure 3. Multifrequency EEG Decoding Reveals Predictive Signals of Trial-to-Trial Time Production (TP), Metacognitive Inference (MI), and Confidence with EEG-ViT.

a Trial selection based on component alignment improves decoding accuracy. Scatter-density plots show average 5-fold test accuracy per participant for models trained on trials ranked by correlation with the accuracy-maximizing principal component (top 30%, 50%, or 100%). P-values indicate pairwise comparisons between trial selection groups (independent two-sample t-tests). Classification accuracy increased with stricter trial selection, with the top 30% yielding the highest performance. One-sample t-tests showed that accuracy in the top 30% and top 50% exceeded chance level (33%), whereas the 100% condition showed only trend-level effects. Accuracy differences were largest between the top 30% and 100%,

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

followed by 30% and 50%, and 50% and 100%. The lower panel shows that trials with stronger component correlations exhibited clearer temporal rebound patterns, emphasizing the importance of selective trial inclusion.

b TP categories are decodable on a single-trial basis before interval completion. Test accuracies for TP category classification (short, accurate, long) are shown per participant using the top 30% of matched trials across all frequency bands. Colored points represent individual cross-validation fold results; black points denote fold-averaged accuracy per participant. Despite fold-to-fold variability, accuracies consistently exceeded chance level (33%; red line). The blue line indicates the grand mean across participants, confirming reliable trial-by-trial prediction of TP category before interval termination.

c Predictive neural signals exhibit frequency-dependent spatial topographies. Topographic maps display the spatial distribution of selected channels under the top 30% matching criterion for theta (θ; 3–7 Hz; green), alpha (α; 8–12 Hz; blue), beta (β; 13–30 Hz; red), and the broadband combination (θ+α+β; 3–30 Hz; grey). β-band channel selections clustered over central and parietal regions, consistent with the established topography of post-movement β rebound. All decoding relied exclusively on neural activity recorded after TP onset and before offset, with predictive information preserved within a 1.2 s post-onset window, indicating that behaviorally relevant signals emerged well before interval completion.

d Multiple frequency bands contribute independently to TP category decoding. Density plots summarize TP category classification accuracies across frequency bands under the top 30% matching condition. Decoding accuracy significantly exceeded chance (33%) for all bands (θ, α, and β), demonstrating that TP prediction was not exclusively dependent on β-rebound activity but also reflected temporally structured oscillations in the θ and α ranges. α-band activity showed the weakest decoding performance, consistent with prior evidence linking α dynamics to general timing variability.

e MI requires cross-frequency integration. Single-trial MI accuracy (alignment between subjective temporal error estimates and actual TP performance) was decoded before interval termination. Significant decoding was observed only when all frequency bands (θ+α+β) were combined, with α contributing at the trend level. Individual frequency bands alone did not support reliable decoding. These findings indicate that metacognitive inference depends on

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the integration of information across multiple oscillatory sources, rather than being localized to a single frequency band.

f Confidence monitoring engages beta and alpha oscillatory mechanisms. Confidence ratings associated with error monitoring were decoded before TP termination. Both α and β signals carried predictive information, with β showing the stronger effect. θ decoding was present only at the trend level. These findings suggest that confidence monitoring engages oscillatory mechanisms linked to motor-action monitoring (β) and attentional control (α).

6.1) Decoding Time Production and Metacognitive Processes from EEG Using ViT-Based Modeling

Influence of Trial Selection on TP Decoding

We first examined whether selectively training the model on trials whose oscillatory dynamics most closely matched the dominant neural pattern would improve TP category decoding. Motor action timing engages rebound dynamics across β, α, and θ frequency bands; however, these signals fluctuate substantially across trials in latency, amplitude, and spatial distribution, meaning that not all trials carry equally reliable behaviorally relevant information. The accuracy-maximizing principal component, defined as the component yielding the highest decoding performance, captured this dominant oscillatory structure across all three bands, with β showing the strongest expression. This is consistent with the well-documented variability of post-movement β rebound timing, with mean onset around 230 ms and peak activity between 500–1000 ms after movement termination[34,35], and single-trial latency spans of ±400 ms within individuals.[36] Analogous rebound patterns were observed in the α and θ bands, though less pronounced, confirming that the component reflected a multi-band rather than purely β-driven signal. Trials were then ranked by their alignment with this component, and decoding was performed separately on the top 30%, top 50%, and full 100% of trials, to test whether prioritizing trials that most strongly expressed this shared oscillatory structure would yield more reliable TP category classification.

Scatter-density plots (Figure 3a) display mean 5-fold cross-validated test accuracy per participant across trial selection conditions. One-sample *t*-tests confirmed that decoding accuracy significantly exceeded chance (33%) in both the top 30% subset (M = .377, 95% CI [.359, .394], p < .001) and the top 50% subset (M = .354, 95% CI [.341, .367], p < .001). Accuracy in the 100% condition (M = .339, 95% CI [.328, .349], p = .092) was numerically

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

above chance but did not reach statistical significance. Independent two-sample *t*-tests revealed that the top 30% subset significantly outperformed both the top 50% ($t(28) = -2.092$, $p = .041$) and 100% subsets ($t(28) = -3.779$, $p < .001$), while the difference between the top 50% and 100% subsets remained marginal ($t(28) = -1.860$, $p = .069$). Trials most strongly correlated with the accuracy-maximizing component displayed clearer β-rebound-like temporal dynamics (Figure 3a, lower panel), suggesting that selective trial inclusion enriches the neural signal available for decoding. When all frequency bands were combined, decoding accuracy reached up to 57%, subject to cross-fold variability (Figure 3b). Centro frontal, central, parietal, and occipital channels were most consistently selected under the accuracy-maximizing component criterion (Figure 3c), in line with topographies typically associated with beta-rebound activity.[4,17,18]

Taken together, these findings indicate that decoding performance benefits from selectively prioritizing trials that strongly express the dominant temporal pattern of the accuracy-maximizing component. Such high-matching trials exhibit a characteristic spatiotemporal signature consistent with β-rebound dynamics, which enhances the extraction of behaviorally relevant neurophysiological features and improves TP category decoding.

Contributions of Frequency Bands to TP Category Decoding

Classification accuracy was compared across frequency bands under the top 30% matching condition (Figure 3d). One-sample *t*-tests confirmed significant above-chance decoding for all frequency bands: θ ($t(28) = 5.188$, $p < .001$, $M = .371$, 95% CI [.355, .388]), α ($t(28) = 2.811$, $p = .009$, $M = .354$, 95% CI [.337, .372]), β ($t(28) = 6.439$, $p < .001$, $M = .377$, 95% CI [.362, .393]), and the broadband combination θ+α+β ($t(28) = 5.401$, $p < .001$, $M = .377$, 95% CI [.359, .394]). A significant difference between α and β ($t(28) = -2.044$, $p = .046$) indicated that β-band activity carried stronger predictive information for TP categorization. A trend-level difference was observed between α and θ+α+β ($t(28) = -1.829$, $p = .073$). No other pairwise comparisons reached significance, suggesting that broadband combination did not substantially improve decoding accuracy beyond β alone.

Decoding MI categories

Having established that TP categories were reliably decodable from individual frequency bands, we next examined whether the same approach could decode metacognitive inference (MI), the trial-by-trial alignment between subjective error reports and objective timing performance (Figure 3e). This distinction is theoretically critical: single-process accounts

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

predict MI should be decodable from the same bands as TP, given shared decision signals, whereas HOR theory predicts that metacognitive inference requires integration across multiple oscillatory sources. MI categories were therefore decoded using the same ViT architecture across θ, α, and β bands individually and in combination, enabling a direct test of whether first- and second-order neural representations draw on the same or distinct frequency-band resources.

Decoding results revealed a clear dissociation from the TP pattern. Significant above-chance classification was observed exclusively for the broadband combination (θ+α+β; M = .535, $t(28) = 2.976$, p = .006, 95% CI [.511, .559]), while no individual band supported reliable decoding: α (M = .519, $t(28) = 1.754$, p = .091), β (M = .508, $t(28) = 0.886$, p = .383), and θ (M = .507, $t(28) = 0.786$, p = .439). Trend-level differences were observed between the broadband combination and both θ ($t(28) = -1.891$, p = .064) and β ($t(28) = -1.888$, p = .064) individually. These results indicate that reliable MI decoding requires simultaneous integration across all three oscillatory sources, a fundamentally different pattern from TP decoding, where individual bands were each sufficient. Consistent with this, the data-driven principal component features captured trial-to-trial stability of each band's spatiotemporal pattern with respect to TP prediction, with β contributing most strongly, followed by α and θ (Figure 3d).

Decoding Confidence Judgments

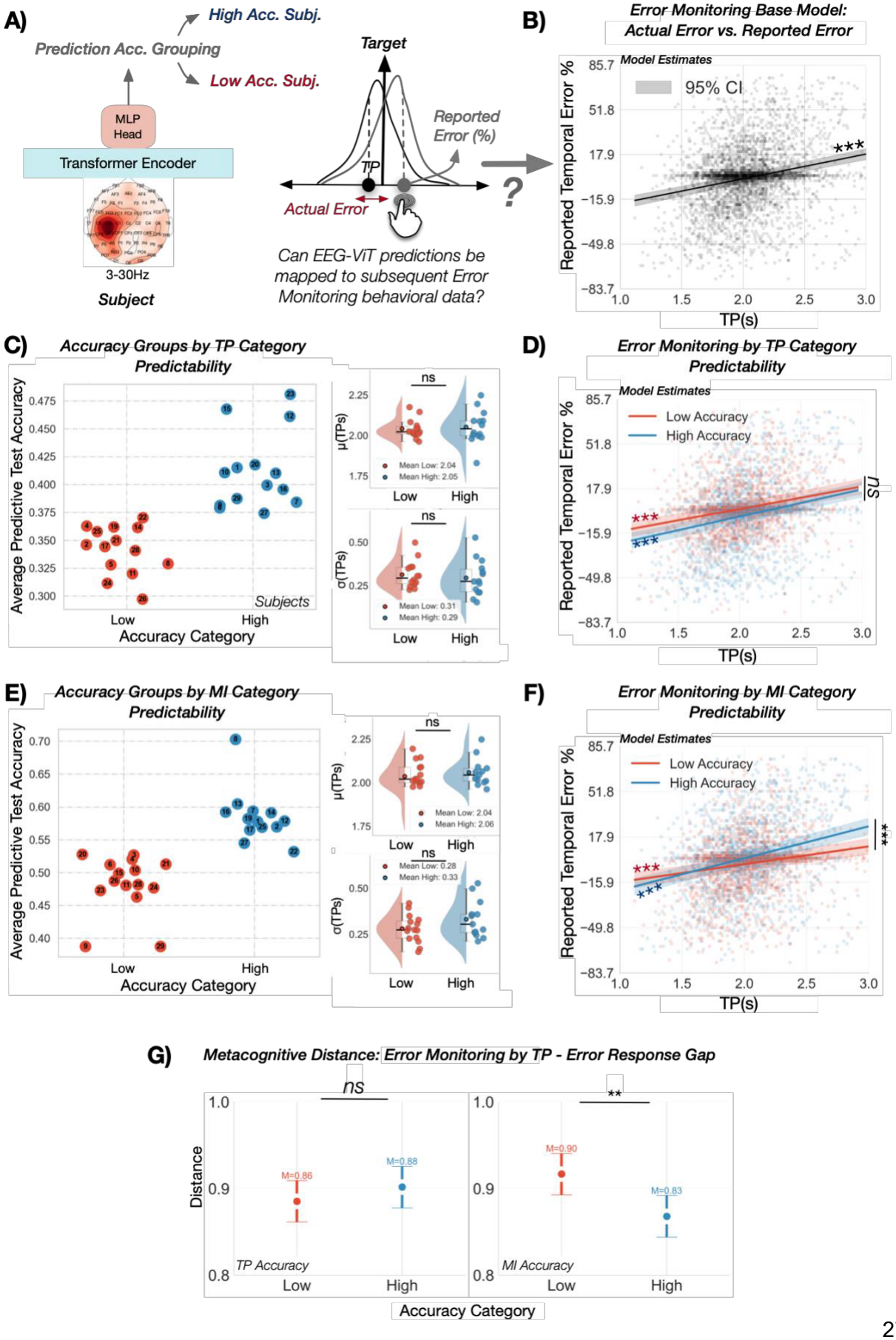
Having dissociated the frequency-band requirements for TP and MI decoding, we next examined whether confidence judgments, participants' subjective certainty about their own error estimates, could be decoded from pre-termination oscillatory activity. Confidence represents a distinct metacognitive signal from MI: whereas MI captures the accuracy of error evaluation, confidence reflects the subjective sense of certainty accompanying it. Decoding confidence from the same pre-termination EEG, therefore, allowed us to ask whether subjective certainty draws on the same oscillatory resources as metacognitive inference, or whether it is supported by a partially separable neural signature.

Confidence ratings deviated significantly from normality across all frequency bands (Shapiro-Wilk: θ, p = .010; α, p = .010; β, p < .001; θ+α+β, p = .010) and were therefore analyzed using nonparametric tests (Figure 3f). Wilcoxon signed-rank tests indicated significant above-chance decoding for β (W = 328, p = .002) and α (W = 284, p = .033), while θ (W = 266, p = .078) and the broadband combination (W = 266, p = .078) reached only trend level. Mann-Whitney U tests revealed no significant differences in decoding accuracy across bands. Unlike MI, which required simultaneous integration across all three bands, confidence

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was primarily supported by α- and β-band dynamics, with β showing the stronger effect, consistent with prior evidence linking β-band activity to post-movement evaluation and internal performance monitoring. This pattern suggests that subjective certainty draws on a partially distinct oscillatory substrate from metacognitive inference itself.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4. Pre-Termination Neural Decoding of Metacognitive States Predicts Behavioral Error Monitoring.

a Decoding-based stratification pipeline linking pre-termination neural dynamics to behavioral error monitoring. Pre-termination refers to neural activity recorded during ongoing motor action timing, before the button press that terminates the produced interval. This activity was decoded using EEG-ViT to classify TP and MI categories. Reported error reflects participants' subjective temporal error estimates converted to percentages, where sign indicates deviation direction (negative = shorter than target; positive = longer) and magnitude quantifies error size. MI category accuracy captures the trial-by-trial alignment between objective timing performance and subjective error reports. Participants were stratified into High and Low decoding accuracy groups based on model classification performance for each target separately, enabling independent tests of two competing predictions: single-process accounts predict that stronger TP neural decoding should correlate with stronger behavioral error monitoring, whereas HOR theory predicts no such relationship for TP decoding but a selective one for MI decoding.

b Trial-level behavioral validation: temporal deviations predict reported error. The relationship between TPs and reported error percentage was examined at the single-trial level using linear mixed-effects models. The solid black line shows the population-level fixed-effect estimate, and the shaded area indicates the 95% confidence interval. Random intercepts modelled participant-level variability across repeated measurements. Scatter points represent individual trial data from all participants. The significant positive correlation indicates that deviations in either direction (shorter or longer TPs) were associated with higher reported error percentages, confirming effective trial-by-trial error monitoring. Statistical significance: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ns $p > 0.05$.

c Participant stratification by TP category decoding accuracy. The strip plot shows individual participants' average pre-termination EEG-ViT classification accuracy for TP categories, stratified into Low (red) and High (blue) TP decoding accuracy groups. Each dot represents one participant, with horizontal jitter applied for visibility and subject identifiers displayed within markers. Insets show group-level timing-response means (top) and standard deviations (bottom), revealing no significant differences in behavioral performance between the Low and High TP decoding accuracy groups.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

d TP category decoding accuracy does not modulate error monitoring strength. Model-implied correlations between TPs and reported error percentage are shown separately for the Low and High TP decoding accuracy groups. Scatter points represent individual trial data for the High (blue) and Low (red) decoding accuracy groups. Both groups exhibited positive correlations with comparable slopes. Contrary to single-process predictions, TP decoding accuracy did not significantly correlate with differences in subsequent behavioral error monitoring strength.

e Participant stratification by MI category decoding accuracy. The strip plot shows individual participants' average pre-termination EEG-ViT classification accuracy for MI categories, stratified into Low (red) and High (blue) MI decoding accuracy groups. Visualization conventions follow panel c. Insets display group-level timing performance means (top) and standard deviations (bottom), indicating no significant differences in behavioral performance between the Low and High MI decoding accuracy groups.

f Higher MI category decoding accuracy selectively predicts stronger behavioral error monitoring. Model-implied correlations between TPs and reported error percentage are shown separately for the Low and High MI decoding accuracy groups. Both groups showed positive correlations, but the correlation was significantly stronger in the High MI decoding group. Consistent with the HOR theory, participants whose pre-termination neural dynamics more reliably decoded MI categories exhibited stronger behavioral error monitoring. This establishes a direct link between neural decoding of metacognitive states and subsequent error monitoring behavior.

g Metacognitive distance is reduced in the high MI category decoding accuracy group. Bar plots show metacognitive distance (MD), computed as the difference between z-scored temporal error responses and z-scored TPs, providing a trial-wise measure of alignment between participants' error monitoring estimates and actual motor timing performance. Error bars indicate the 95% confidence interval of the group mean. MI categories were dichotomized using a median split of MD values across all trials: trials below the median were classified as High MI (more accurate metacognitive inference) and trials at or above the median as Low MI, on the assumption that smaller differences reflect more accurate metacognitive inference. MD was significantly lower in the High MI decoding accuracy group than in the Low MI group, but

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

only when participants were stratified by MI decoding accuracy and not by TP decoding accuracy. Together, these findings support the higher-order representation (HOR) account: more accurate pre-termination neural decoding of metacognitive states predicted superior subsequent metacognitive performance, reflected in a closer alignment between objective performance and reported error percentage.

6.2) Neural predictability of metacognitive states maps onto behavioral error monitoring.

Two competing theoretical accounts generate opposing predictions about the relationship between pre-termination neural activity and subsequent error monitoring behavior, yet direct neural evidence adjudicating between them has been lacking. Single-process accounts propose that first-order performance and metacognitive evaluation arise from a shared decision variable, predicting that participants whose pre-termination neural activity more accurately encodes timing performance should also exhibit stronger error monitoring.[12,16] HOR theory, by contrast, predicts a selective dissociation: no such relationship for TP decoding accuracy, but a specific link between MI decoding accuracy and error monitoring strength.[14,15] To provide this first direct test, we examined whether inter-individual differences in pre-termination neural decoding accuracy, separately for TP and MI categories, predicted subsequent behavioral error monitoring (Figure 4a). Two decoding targets were distinguished: TP category decoding accuracy, reflecting the extent to which pre-termination neural activity predicted first-order timing outcomes, and MI category decoding accuracy, reflecting the extent to which it predicted the trial-by-trial alignment between objective performance and subjective error reports.

Before testing these theoretical predictions, we confirmed reliable trial-level behavioral error monitoring. Produced intervals were significantly correlated with reported error percentage (b = 18.69, SE = 1.25, z = 14.97, p < .001, 95% CI [16.24, 21.13]; Figure 4b), confirming that subjective error evaluations systematically tracked actual performance on a trial-by-trial basis.

To test single-process predictions, participants were stratified into Low and High TP decoding accuracy groups (Figure 4c), which did not differ in mean timing accuracy (t = −0.34, p = .737) or variability (t = 0.54, p = .593). Despite robust inter-individual differences in TP decoding accuracy, the correlation between produced intervals and reported error was comparable across groups, with no significant interaction (b = 3.54, SE = 2.51, z = 1.41, p =

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

.158, 95% CI [−1.37, 8.46]; Figure 4d). Including decoding accuracy group as an interaction term did not improve model fit (ΔAIC = −0.02; $\chi^2(1) = 2.02$, $p = .155$), inconsistent with single-process predictions.

MI decoding accuracy, by contrast, was selectively associated with behavioral error monitoring, consistent with HOR theory. Participants stratified into Low and High MI decoding accuracy groups (Figure 4e) did not differ in mean timing accuracy ($t = −0.70$, $p = .488$) or variability ($t = −1.52$, $p = .141$), ruling out performance-level confounds. Although both groups showed positive correlations between produced intervals and reported error, this relationship was significantly stronger in the High MI decoding accuracy group ($b = 10.61$, $SE = 2.51$, $z = 4.24$, $p < .001$, 95% CI [5.70, 15.52]; Figure 4f), with the inclusion of decoding accuracy group as an interaction term substantially improving model fit (ΔAIC = −15.83; $\chi^2(1) = 17.83$, $p < .001$).

This dissociation was further validated using metacognitive distance (MD), defined as the absolute difference between z-scored reported error and z-scored produced intervals (Figure 4g). Stratification by TP decoding accuracy yielded no group difference in MD ($t = −0.949$, $p = .343$; Low: $M = 0.857$, 95% CI [0.827, 0.887]; High: $M = 0.877$, 95% CI [0.847, 0.907]). Stratification by MI decoding accuracy, however, revealed significantly lower MD in the High group ($t = 2.809$, $p = .005$; Low: $M = 0.896$, 95% CI [0.866, 0.926]; High: $M = 0.835$, 95% CI [0.805, 0.865]), reflecting superior alignment between objective performance and subjective error estimates.

Together, these findings reveal a functional dissociation in the behavioral relevance of pre-termination oscillatory activity: TP decoding accuracy did not modulate subsequent error evaluation, whereas MI decoding accuracy predicted both stronger performance–error correlations and reduced metacognitive distance (Figures 4f and 4g).

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

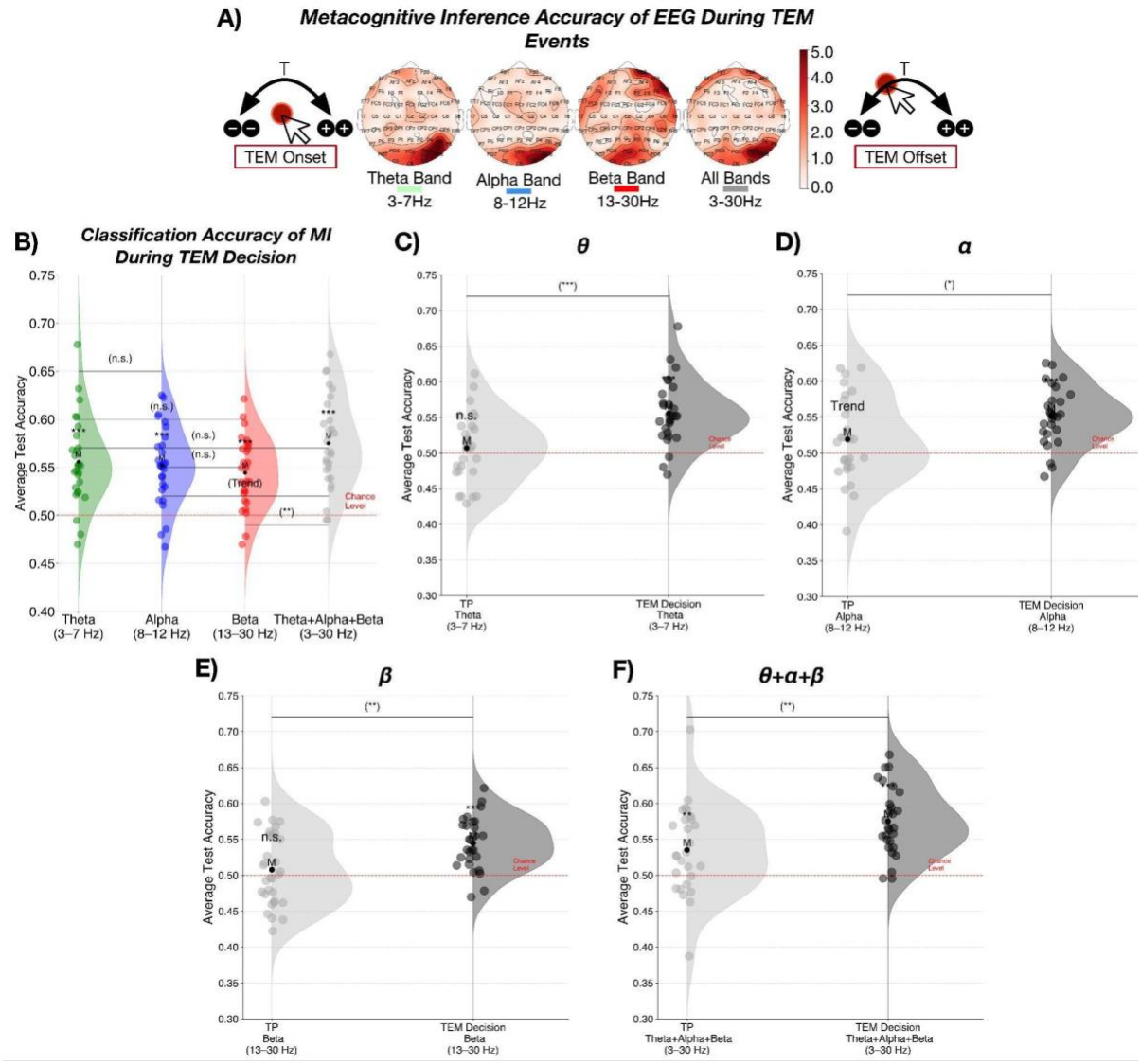


Figure 5. Distributed Oscillatory Signals Track MI During the TEM Decision Phase.

a Parieto-frontal topography of MI category decoding during error evaluation. Trials were selected based on the strength of correlation with each participant's accuracy-maximizing principal component, corresponding to MI test accuracy. EEG activity was measured during the temporal error monitoring period (between response onset and offset). Channel selection frequency peaked at POz, P6, PO4, P4, P8, PO8, AFz, AF8, CP6, and CP2, revealing distributed parieto-frontal contributions to single-trial MI decoding.

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

b Cross-frequency integration maximizes MI decoding accuracy. Each dot represents participant-wise, fold-averaged test accuracy. Decoding exceeded chance level (50%) for all frequency bands combined, as well as for θ, α, and β individually. Combined-band accuracy was higher than that of α or β alone, but not significantly higher than θ, suggesting a prominent contribution of θ-band activity to metacognitive processing and indicating that integrating information across frequency bands maximizes the extraction of higher-order evaluative signals.

c–f Temporal dynamics of metacognitive signals across task phases. Classification accuracy is compared between the TP onset (light gray) and the TEM decision phase (dark gray). Accuracy was higher during the TEM decision phase across all frequency bands, indicating that metacognitive signals strengthen during explicit error evaluation. However, combined-band activity showed above-chance MI decoding already at TP onset, suggesting that evaluative neural markers are present before explicit decision-making. α-band activity exhibited a trend-level effect at TP onset, indicating that specific oscillatory components may contribute to early metacognitive evaluation. Overall, these findings show that metacognitive signals peak during the decision phase but remain partially detectable during ongoing action timing, suggesting that TEM and higher-order evaluation unfold continuously rather than being confined to a discrete decision point.

6.3) Metacognitive inference was strongly decoded from distributed EEG activity during the TEM decision phase.

The preceding analyses established that pre-termination oscillatory dynamics selectively predicted behavioral error monitoring. A remaining question is whether metacognitive signals were specific to the action execution phase or whether they strengthened further as participants transitioned into explicit error evaluation. To address this, we compared MI decoding accuracy between the TP onset and TEM decision phases across all frequency bands, allowing us to characterize both the temporal evolution of metacognitive neural signals and the consistency of their frequency-band requirements across task stages.

Classification accuracy during the TEM decision phase demonstrated that MI could be reliably decoded from all frequency bands: *θ* ($t(28) = 6.548$, $p < .001$, $M = 0.556$, 95% CI [0.538, 0.573]), *α* ($t(28) = 6.763$, $p < .001$, $M = 0.552$, 95% CI [0.536, 0.567]), *β* ($t(28) = 6.398$, $p < .001$, $M = 0.544$, 95% CI [0.530, 0.558]), and the broadband combination θ+α+β ($t(28) = 8.434$,

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$p < .001$, $M = 0.575$, 95% CI [0.557, 0.593]; Figure 5b). Combined-band decoding significantly exceeded both *α* alone ($t(28) = -1.980$, $p < .05$) and *β* alone ($t(28) = -2.713$, $p < .01$), but not *θ* alone ($t(28) = -1.559$, $p = .091$), suggesting that *θ*-band activity plays a particularly prominent role in metacognitive processing and that cross-frequency integration provides maximal predictive information.

Comparisons between TP onset and the TEM decision phase revealed clear temporal specificity in the emergence of MI-related signals (Figure 5c–f). At TP onset, *θ*- and *β*-band decoding did not exceed chance, but both became significant during the TEM decision phase (*θ*: $t(28) = -3.911$, $p < .001$; *β*: $t(28) = -3.221$, $p = .002$). *α*-band decoding showed only a trend-level effect at TP onset but reached significance during the decision phase ($t(28) = -2.429$, $p = .019$). Notably, combined-band decoding (θ+α+β) was already above chance at TP onset and increased further during the TEM decision phase ($t(28) = -2.707$, $p = .009$), indicating that cross-frequency integration provided the strongest and most sustained metacognitive signal across both task phases.

Overall, these results indicate that although MI-related neural signals were detectable during action execution, their predictive strength peaked during the TEM decision phase. This pattern of temporally graded emergence is consistent with a continuously accumulating higher-order evaluative process, in which metacognitive information becomes maximally accessible immediately preceding explicit error judgment.

7) Discussion

The present study investigated whether metacognitive evaluation of motor action timing can be explained by the same neural signals that support timing performance, or whether it reflects at least partially distinct processes involving additional evaluative computations. To adjudicate between these accounts, we applied Vision Transformer decoding to single-trial EEG recorded across θ, α, and β frequency bands during both the motor action execution and temporal error monitoring (TEM) phases. The findings converge on a consistent picture: motor action timing and metacognitive evaluation are supported by partially dissociable neural architectures, with the latter requiring a form of multi-band integration that the former does not.

Time production categories were reliably decodable from any individual frequency band. θ, α, and β each contributed sufficient information for classification, and broadband combination yielded no substantial improvement. This pattern of partially redundant, band-sufficient encoding is consistent with prior evidence that β-band oscillations index temporal

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

intervals at the single-trial level,[17,18] and propose that frequency-specific oscillatory codes provide parallel, largely independent representations of interval duration. The selective contribution of β and θ to TP decoding, and the relative weakness of α, aligns with the established topography of post-movement β rebound[4,17,18] and with fronto-parietal θ coordination during motor action and decision processes.[37,38]

In contrast, metacognitive inference (MI), the trial-by-trial alignment between objective timing performance and subjective error evaluation, was decodable only when θ, α, and β were integrated simultaneously. No individual band supported reliable MI classification. This dissociation is not merely a quantitative difference in decoding strength; it reflects a qualitative difference in representational architecture. Single-process accounts[12,16] predict that metacognitive evaluation should be decodable from the same signals as timing performance, since both arise from a common accumulator variable. The data directly contradict this prediction. The failure of individual bands to support MI decoding, combined with the success of their joint integration, implies that metacognitive evaluation draws on coordinated multi-band dynamics that are not reducible to any single oscillatory substrate. This is consistent with the Higher-Order Representation (HOR) framework,[14,15] which holds that conscious awareness of a mental state requires a higher-order re-representation that takes the first-order state as its object rather than simply reflecting it.

The role of β within this multi-band context deserves particular attention. β participated in MI decoding only as part of the integrated broadband signal, not as an independent predictor, a pattern distinct from its role in TP decoding, where it was sufficient alone. This dissociation is consistent with proposals that β reflects the maintenance and reactivation of internal representational states beyond simple motor execution feedback,[39,40] and with direct evidence that post-movement β indexes confidence in internal model estimates in sensorimotor contexts.[41] The present findings extend this interpretation to the domain of temporal metacognition: β contributes to metacognitive evaluation not as a motor readout but as part of a coordinated higher-order representation that also requires θ- and α-band contributions.

Importantly, MI was already decodable during ongoing motor execution, up to 1 s before action termination, and this decoding was strengthened further during the explicit TEM decision phase. This temporal profile indicates that evaluative information accumulates continuously during interval production rather than arising solely at the point of explicit judgment. This finding aligns with Bayesian accounts of temporal uncertainty,[42,43] under which uncertainty about the current production is represented as a structured signal that accumulates progressively during

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

interval generation and is available for metacognitive evaluation prior to any explicit response. It is also consistent with continuous evidence accumulation frameworks in which evaluative signals build progressively in parallel with ongoing performance, rather than emerging only after task completion.[44,45]

Spatial evidence further supported this dissociation. TP decoding localized to centro-parietal and frontal-central regions, consistent with the established topography of motor action timing and post-movement β rebound. MI decoding, by contrast, recruited a distributed parieto-frontal network spanning posterior and prefrontal sites, consistent with the integration of temporal performance representations and prefrontal evaluative processes.[46] This spatial shift from execution-related to evaluative networks is consistent with neural evidence that dedicated prefrontal circuits contribute to higher-order representations in ways that cannot be reduced to first-order activity.[3,15]

The most direct evidence against single-process accounts came from the neural–behavioral dissociation. Participants stratified by TP decoding accuracy showed no difference in error monitoring strength or metacognitive distance, directly contradicting the prediction that stronger timing encoding should produce more accurate self-evaluation. Participants stratified by MI decoding accuracy, by contrast, showed significantly stronger performance–error response correlations and reduced metacognitive distance, despite equivalent timing accuracy. Because timing performance was equated across groups, the metacognitive benefit cannot be attributed to superior timing decoding; it must arise from a representational layer that is neurally distinct from performance control. This selective pattern challenges hierarchical accounts as much as single-process ones: even a model that posits a distinct metacognitive stage but ties its precision to the quality of first-order (here, motor timing signals) encoding[13] would predict some relationship between TP decoding accuracy and TEM precision, an absent relationship. The data are instead most consistent with HOR theory,[14,15] under which metacognitive evaluation constitutes an active re-representation that draws on first-order timing states but engages at least partially distinct mechanisms beyond those supporting task execution.

TEM confidence judgments showed a pattern distinct from both TP and MI decoding. Significantly above-chance decoding was observed for α and β individually, without requiring multi-band integration. This suggests that subjective certainty may partly reflect first-order signal properties, β-indexed performance quality and α-indexed attentional certainty, rather than the higher-order inferential process that underlies accurate metacognitive evaluation. The dissociation between confidence and MI decoding raises the possibility that these two aspects

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of metacognition, the accuracy of self-evaluation and the subjective sense of certainty accompanying it, are at least partially separable at the neural level, a distinction that aligns with signal detection frameworks separating metacognitive sensitivity from response bias.[47,48]

Taken together, these findings establish multi-band neural integration as a necessary condition for temporal metacognition. The dissociation between single-band sufficiency for timing performance and multi-band dependence for metacognitive inference, combined with the selective behavioral relevance of MI decoding accuracy, converges on the conclusion that metacognitive evaluation is not a passive reflection of interval production signals but an active, higher-order process that recruits a representational organization qualitatively distinct from the one supporting primary task performance. Within the broader landscape of temporal metacognition,[4] this constitutes the first single-trial neural evidence directly distinguishing first-order timing representations from second-order metacognitive ones, and provides direct support for the view that self-evaluative processing requires a dedicated neural architecture consistent with HOR theory.

8) Limitations and Future Directions

Several methodological considerations merit acknowledgement. First, trial–channel observations per participant after preprocessing were modest. Replication with larger trial samples and multi-session designs may represent an important future step. Future work might also benefit from data augmentation techniques to expand single-trial training sets,[49] and from study designs incorporating subject-specific EEG architectures to better accommodate inter-participant variability.[50]

Second, isolating metacognitive signals at the single-trial level is inherently challenging: the brain may not always produce clearly detectable β-rebound-like patterns, and metacognitive signals may be particularly difficult to decode reliably. To address this, we developed a PCA-based pipeline retaining the top 30% of oscillatory-structure-aligned trials, a step that may have been critical for reliable decoding and could offer a useful methodological contribution for future single-trial metacognitive EEG studies.

Third, recording under a single motor condition leaves open whether the θ–α–β integration pattern identified here might be disrupted or reorganized under higher motor demands. Given that increased motor demands selectively impair TEM precision without affecting timing accuracy,[22] future work combining parametric motor manipulations with single-

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

trial EEG decoding could potentially provide the mechanistic bridge needed to connect the behavioral and neural findings of this thesis.

9) Code and Data Availability

The data and code supporting the findings of this study will be made openly available on GitHub at https://github.com/TEM-Opus2020 upon publication.

10) Funding Declaration

This study was supported thanks to the OPUS grant 2019/35/B/HS6/04389 to TWK from the National Center of Science in Poland.

11) References

1. Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

2. Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

3. Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B, 367*(1594), 1338–1349. https://doi.org/10.1098/rstb.2011.0417

4. Kononowicz, T. W., Roger, C., & van Wassenhove, V. (2019). Temporal metacognition as the decoding of self-generated brain dynamics. *Cerebral Cortex, 29*(10), 4366–4380. https://doi.org/10.1093/cercor/bhy318

5. Bilgin, S. N., Doyère, V., & Kononowicz, T. W. (2026). Direct comparison of temporal error monitoring in humans and rats. *iScience, 29*(1), 115531. https://doi.org/10.1016/j.isci.2026.115531

6. Bilgin, S. N., & Kononowicz, T. W. (2025). Temporal error monitoring: Monitoring of internal clock or just motor noise? *Consciousness and Cognition, 130*, 103849. https://doi.org/10.1016/j.concog.2025.103849

7. Akdoğan, B., & Balcı, F. (2017). Are you early or late?: Temporal error monitoring. *Journal of Experimental Psychology: General, 146*(3), 347–361. https://doi.org/10.1037/xge0000265

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

8. Kononowicz, T. W., & van Wassenhove, V. (2019). Evaluation of self-generated behavior: Untangling metacognitive readout and error detection. *Journal of Cognitive Neuroscience, 31*(11), 1641–1657. https://doi.org/10.1162/jocn_a_01442

9. Öztel, T., & Balcı, F. (2023). Humans can monitor trial-based but not global timing errors: Evidence for relative judgements in temporal error monitoring. *Quarterly Journal of Experimental Psychology, 76*(9), 2155–2163. https://doi.org/10.1177/17470218221145314

10. Öztel, T., & Balcı, F. (2024). Metric error monitoring as a component of metacognitive processing. *The European Journal of Neuroscience, 59*(5), 807–821. https://doi.org/10.1111/ejn.16182

11. Kononowicz, T. W., van Wassenhove, V., & Doyère, V. (2022). Rodents monitor their error in self-generated duration on a single-trial basis. *Proceedings of the National Academy of Sciences, 119*(9), e2108850119. https://doi.org/10.1073/pnas.2108850119

12. Balcı, F., & Simen, P. (2016). A decision model of timing. *Current Opinion in Behavioral Sciences, 8*, 94–101. https://doi.org/10.1016/j.cobeha.2016.02.002

13. Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124*(1), 91–114. https://doi.org/10.1037/rev0000045

14. Rosenthal, D. M. (1986). Two concepts of consciousness. *Philosophical Studies, 49*(3), 329–359. https://doi.org/10.1007/BF00355521

15. Brown, R., Lau, H., & LeDoux, J. E. (2019). Understanding the higher-order approach to consciousness. *Trends in Cognitive Sciences, 23*(9), 754–768. https://doi.org/10.1016/j.tics.2019.06.009

16. Balcı, F., & Simen, P. (2014). Decision processes in temporal discrimination. *Acta Psychologica, 149*, 157–168. https://doi.org/10.1016/j.actpsy.2014.03.005

17. Kononowicz, T. W., & van Rijn, H. (2015). Single-trial beta oscillations index time estimation. *Neuropsychologia, 75*, 381–389. https://doi.org/10.1016/j.neuropsychologia.2015.06.014

18. Kulashekhar, S., Pekkola, J., Palva, J. M., & Palva, S. (2016). The role of cortical beta oscillations in time estimation. *Human Brain Mapping, 37*(9), 3262–3281. https://doi.org/10.1002/hbm.23239

19. Mioni, G., Shelp, A., Stanfield-Wiswell, C. T., Gladhill, K. A., Bader, F., & Wiener, M. (2020). Modulation of individual alpha frequency with tACS shifts time perception. *Cerebral Cortex Communications, 1*(1), tgaa064. https://doi.org/10.1093/texcom/tgaa064

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

20. Rohenkohl, G., & Nobre, A. C. (2011). Alpha oscillations related to anticipatory attention follow temporal expectations. *Journal of Neuroscience, 31*(40), 14076–14084. https://doi.org/10.1523/JNEUROSCI.3387-11.2011

21. Barne, L. C., Claessens, P. M. E., Reyes, M. B., Caetano, M. S., & Cravo, A. M. (2017). Low-frequency cortical oscillations are modulated by temporal prediction and temporal error coding. *NeuroImage, 146*, 40–46. https://doi.org/10.1016/j.neuroimage.2016.11.028

22. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv*. https://doi.org/10.48550/arXiv.2010.11929

23. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

24. World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, *310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053

25. Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

26. Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference* (pp. 57–61). https://doi.org/10.25080/Majora-92bf1922-011

27. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

28. Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460. https://doi.org/10.1016/j.neuroimage.2013.10.027

29. Larson, E., Gramfort, A., Engemann, D. A., Leppakangas, J., Brodbeck, C., Jas, M., Brooks, T. L., Sassenhagen, J., McCloy, D., Luessi, M., King, J.-R., Höchenberger, R., Goj, R., Brunner, C., Favelier, G., van Vliet, M., Wronkiewicz, M., Rockhill, A., & Holdgraf, C. (2024). *MNE-Python* (Version 1.9.0) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.14519545

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

30. Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, *159*, 417–429. https://doi.org/10.1016/j.neuroimage.2017.06.030

31. Delorme, A. (2023). EEG is better left alone. *Scientific Reports*, *13*(1), 2372. https://doi.org/10.1038/s41598-023-27528-0

32. Szeghalmy, S., & Fazekas, A. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. *Sensors*, *23*(4), 2333. https://doi.org/10.3390/s23042333

33. Schroeder, F., Fairclough, S., Dehais, F., & Richins, M. (2025). The impact of cross-validation choices on pBCI classification metrics: Lessons for transparent reporting. *Frontiers in Neuroergonomics*, *6*, 1582724. https://doi.org/10.3389/fnrgo.2025.1582724

34. Bailey, L. M., & Bardouille, T. (2025). Demonstrating the need for long inter-stimulus intervals when studying the post-movement beta rebound following a simple button press. *Frontiers in Neuroscience*, *19*, 1547916. https://doi.org/10.3389/fnins.2025.1547916

35. Jurkiewicz, M. T., Gaetz, W. C., Bostan, A. C., & Cheyne, D. (2006). Post-movement beta rebound is generated in motor cortex: Evidence from neuromagnetic recordings. *NeuroImage*, *32*(3), 1281–1289. https://doi.org/10.1016/j.neuroimage.2006.06.005

36. Little, S., Bonaiuto, J., Barnes, G., & Bestmann, S. (2019). Human motor cortical beta bursts relate to movement planning and response errors. *PLOS Biology*, *17*(10), e3000479. https://doi.org/10.1371/journal.pbio.3000479

37. Rawle, C. J., Miall, R. C., & Praamstra, P. (2012). Frontoparietal theta activity supports behavioral decisions in movement-target selection. *Frontiers in Human Neuroscience, 6*, 138. https://doi.org/10.3389/fnhum.2012.00138

38. Praamstra, P., Kourtis, D., & Nazarpour, K. (2009). Simultaneous preparation of multiple potential movements: Opposing effects of spatial proximity mediated by premotor and parietal cortex. *Journal of Neurophysiology, 102*(4), 2084–2095. https://doi.org/10.1152/jn.00413.2009

39. Engel, A. K., & Fries, P. (2010). Beta-band oscillations — signalling the status quo? *Current Opinion in Neurobiology, 20*(2), 156–165. https://doi.org/10.1016/j.conb.2010.02.015

40. Spitzer, B., & Haegens, S. (2017). Beyond the status quo: A role for beta oscillations in endogenous content (re)activation. *eNeuro, 4*(4), ENEURO.0170-17.2017. https://doi.org/10.1523/ENEURO.0170-17.2017

41. Tan, H., Wade, C., & Brown, P. (2016). Post-movement beta activity in sensorimotor cortex indexes confidence in the estimations from internal models. *Journal of Neuroscience, 36*(5), 1516–1528. https://doi.org/10.1523/JNEUROSCI.3204-15.2016

bioRxiv preprint doi: https://doi.org/10.64898/2026.03.24.714009; this version posted May 19, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

42. Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience, 13*(8), 1020–1026. https://doi.org/10.1038/nn.2590

43. Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences, 17*(11), 556–564. https://doi.org/10.1016/j.tics.2013.09.009

44. Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in decision-making. *Nature, 461*(7261), 263–266. https://doi.org/10.1038/nature08275

45. van den Berg, R., Anandalingam, K., Zylberberg, A., Kiani, R., Shadlen, M. N., & Wolpert, D. M. (2016). A common mechanism underlies changes of mind about decisions and confidence. *eLife, 5*, e12192. https://doi.org/10.7554/eLife.12192

46. Wokke, M. E., Cleeremans, A., & Ridderinkhof, K. R. (2017). Sure I'm sure: Prefrontal oscillations support metacognitive monitoring of decision making. *Journal of Neuroscience, 37*(4), 781–789. https://doi.org/10.1523/JNEUROSCI.1612-16.2016

47. Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 443. https://doi.org/10.3389/fnhum.2014.00443

48. Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition, 21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

49. Lashgari, E., Liang, D., & Maoz, U. (2020). Data augmentation for deep-learning-based electroencephalography. *Journal of Neuroscience Methods*, *346*, 108885. https://doi.org/10.1016/j.jneumeth.2020.108885

50. Perez-Velasco, S., Santamaria-Vazquez, E., Martinez-Cagigal, V., Marcos-Martinez, D., & Hornero, R. (2022). EEGSym: Overcoming inter-subject variability in motor imagery-based BCIs with deep learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *30*, 1766–1775. https://doi.org/10.1109/TNSRE.2022.3186442